Educative Interventions to Combat Misinformation:
Evidence From a Field Experiment in India*

Sumitra Badrinathan†
University of Pennsylvania

May 8, 2020

Abstract

Misinformation makes democratic governance harder, especially in developing countries. Despite its real-world import, little is known about how to combat fake news outside of the U.S., particularly in places with low education, accelerating Internet access, and encrypted information sharing. This study uses a field experiment in India to test the efficacy of a pedagogical intervention on respondents' ability to identify fake news during the 2019 elections (N=1224). Treated respondents received in-person media literacy training in which enumerators demonstrated two tools to identify fake news: reverse image searching and navigating a fact-checking website. Receiving this hour-long media literacy intervention did not significantly increase respondents' ability to identify fake news on average. However, treated respondents who support the ruling party became significantly less able to identify pro-attitudinal fake news. These findings point to the resilience of misinformation in India and the presence of motivated reasoning in a traditionally non-ideological party system.

Keywords: Misinformation, India, Elections, Social Media, Fact-Checking, Literacy Training, WhatsApp

*This study was preregistered with Evidence in Governance and Policy (20190916AA) and received IRB approval from the University of Pennsylvania (832833). The author thanks Devesh Kapur, Guy Grossman, Neelanjan Sircar, Marc Meredith, Matthew Levendusky, Yphtach Lelkes, Milan Vaishnav, Adam Ziegfeld, Devra Moehler, Jeremy Springman, Emmerich Davies, Simon Chauchard and Simone Dietrich. Pranav Chaudhary and Sunai Consultancy provided excellent on-ground and implementation assistance. This research was funded by the Center for the Advanced Study of India (CASI) at the University of Pennsylvania and the Judith Rodin Fellowship. For comments and feedback, the author thanks seminar participants at the NYU Experimental Social Sciences Conference, MIT GOV/LAB Political Behavior of Development Conference, Penn Development and Research Initiative, Southern Political Science Association Meeting, and the Harvard Experimental Political Science Conference.

†PhD Candidate in Political Science. Email: sumitra@sas.upenn.edu

1 Introduction

Images of mutilated bodies and lifeless children proliferated across WhatsApp in northern India in 2018, allegedly resulting from an organized kidnapping network. In response to these messages, a young man mistaken for one of the kidnappers was mobbed and brutally beaten by villagers in Meerut, Haryana. The images, however, were not from a kidnapping network, but rather from a chemical weapons attack in Ghouta, Syria in 2013. Mob lynchings such as this have become a prominent problem in India since 2015, when a Muslim villager in Uttar Pradesh was killed by a mob after rumors spread that he was storing beef in his house. Such misinformation campaigns are often developed and run by political parties with nationwide cyber-armies, targeting political opponents, religious minorities and dissenting individuals (Poonam and Bansal 2019). The consequences of such misinformation are as extreme as violence, demonstrating that fake news is a matter of life and death in India and other developing countries.

What tools, if any, exist to combat the misinformation problem in developing countries? Much of the literature on fact-checking focuses on minor technocratic interventions to reduce misinformation. These include providing warning labels on articles, attaching disputed tags to misinformation, and providing corrective information beside fake claims (Bode and Vraga 2015; Pennycook, Cannon, and Rand 2018; Clayton et al. 2019; Nyhan et al. 2019). But nearly all of the extant literature focuses on the U.S. and other developed democracies, where misinformation spreads via public sites such as Facebook and Twitter. Such interventions are not easily adapted for misinformation distributed on encrypted chat applications such as WhatsApp. On such applications no one, including the app developers themselves, can see, read or analyze messages, hence top-down interventions such as including warning labels beside misinformation are effectively futile. Encryption necessitates that the burden of fact-checking falls solely on the user and therefore, the more appropriate solutions in such contexts are bottom-up, user-driven learning and fact-checking to combat fake news.

This study is one such bottom-up effort to counter fake news with a broad pedagogical program. I investigate whether improving information processing skills changes actual information processing in a partisan environment. The specific research question asked in this paper is whether in-person, pedagogical training to verify information is effective in combating fake news in India. To answer this question, I implemented a large-scale field experiment with 1,224 respondents in the state of Bihar in northern India during the 2019 general elections, when misinformation was arguably at its peak. In an hour-long intervention, treatment group respondents were taught two concrete tools to verify information: performing reverse image searches, and navigating a fact-checking website. They also received a flyer with tips to spot fake news, along with corrections to four political fake stories. After a two-week period, respondent households were revisited to measure their ability to identify fake news.

My experiment shows that an hour-long, educative treatment is not sufficient to help respondents combat fake news. Importantly, the average treatment effect is not significantly distinguishable from zero. Finding that an in-person, hour-long and bottom-up learning inter-

vention does not move people's prior attitudes on fake news is testimony to the tenacity and destructive effects of misinformation in low education settings such as India. It challenges conventional findings in American politics that subtle priming treatments, such as disputed tags, can reduce fake news consumption. These findings also confirm qualitative evidence about the distinctive nature of social media consumers in developing states who are new to the Internet, lending them particularly rife and vulnerable to misinformation.

While there is no evidence of a non-zero average treatment effect, there are significant treatment effects among some subgroups. Bharatiya Janata Party (BJP) partisans (those self-identifying as supporters of the BJP, the national right-wing party in India) who receive the treatment are *less* likely to identify pro-attitudinal stories as fake. That is, on receiving counter-attitudinal corrections, the treatment backfires for BJP respondents while simultaneously working to improve information processing for non-BJP respondents. This is consistent with findings in American politics on motivated reasoning, demonstrating that respondents seek out information reinforcing prior beliefs, and that partisans cheerlead for their party and are likely to respond expressively to partisan questions (Taber and Lodge 2006; Gerber and Huber 2009; Prior, Sood, and Khanna 2015). These findings also challenge the contention that Indians lack consolidated, strong partisan identities (Chhibber and Verma 2018). I demonstrate that party identity in India is more polarized than previously thought, particularly with BJP partisans and during elections.

This is the first analysis of the effect of educative and persuasive interventions on misinformation in India, the world's largest democracy. This study hopes to spark a research agenda on the ways to create an informed citizenry in low income democracies. In doing so, this study hopes to encourage the testing and implementation of bottom-up measures to fight misinformation that are especially suited to settings where information is consumed over encrypted mediums such as WhatsApp. I also seek to contribute to the empirical study of partisan identity in India, revisiting the conventional wisdom of party identities being unconsolidated and fluctuating.

2 What is Fake News and How Do We Fight it?

Following Allcott and Gentzkow (2017), I define "fake news" as stories that are intentionally and verifiably false and could mislead readers. While fake news is now a global phenomenon, the literature on it is grounded in the American context. The predominant model of misinformation in this literature comes from Gentzkow, Shapiro, and Stone (2015). They posit that consumption of misinformation is a result of preferences for confirmatory stories rather than the truth because of the psychological utility from such stories, thereby producing an equilibrium where news outlets are incentivized to report in a biased way. Consumers must make a choice between deriving psychological utility from ideologically-consistent news, or receiving utility from knowing the true state of the world.

The empirical literature on misinformation shows that misperceptions are wide-spread.

Flynn (2016) posits that over 20% of Americans believe misinformation about political issues; Allcott and Gentzkow (2017) estimate that every American adult read about 3 fake stories during the 2016 election. Examples of this include widespread beliefs that certain vaccines can cause autism in healthy children, or that President Obama was not born in the U.S., both of which are demonstrably false (Freed et al. 2010; Nyhan and Reifler 2012). Such misinformed beliefs are especially troubling when they lead people to action, as these skewed views may well alter political behavior (Hochschild and Einstein 2015).

A large research agenda has tested interventions to reduce the consumption of misinformation. Such studies typically rely on top-down interventions, or those emanating from social media platforms themselves. Examples include providing corrections, warnings, or fact-checking and consequently measuring respondents' perceived accuracy of news stories. For instance, in 2016 Facebook began adding "disputed" tags to stories in its newsfeed that had been previously debunked by fact-checkers (Mosseri 2017); it then switched to providing fact checks underneath suspect stories (Smith, Jackson, and Raj 2017). The prevalence for piloting such technocratic solutions to the misinformation problem has since grown rapidly: Chan et al. (2017) find that explicit warnings can reduce the effects of misinformation; Pennycook, Cannon, and Rand (2018) test and find that disputed tags alongside veracity tags can lead to reductions in perceived accuracy; Fridkin, Kenney, and Wintersieck (2015) demonstrate that corrections from professional fact-checkers are more successful at reducing misperceptions.

However, such solutions are limited in their ability to solve problems associated with misinformation. Fact-checking and warning treatments are only effective when misinformation is not salient, when priors are not strong, and when outcomes are measured immediately after intervention, leading to the most significant misperceptions being stable and persistent over time (Nyhan and Reifler 2012; Flynn, Nyhan, and Reifler 2017). Such persistence stems from respondents' motivated reasoning. Human reasoning is pulled between the Scylla of accuracy and the Charybdis of confirmation: humans want to come to the "right" answer, but the architecture of our minds biases us toward information that reinforces our prior beliefs (Kunda 1990). As a result, consistent with the Gentzkow, Shapiro, and Stone (2015) model, we tend to seek out information that reinforces our preferences, counter-argue information that contradicts preferences, and view pro-attitudinal information as more convincing than counter-attitudinal information (Taber and Lodge 2006).

Not surprisingly, then, the large number and comprehensive set of variations in fact-checking and warning treatments has had little success in combating misinformation over time. Further, these studies are almost all lab and survey experiments, and hence we know little about their real-world ecological validity. Empirical findings from studies on media effects that use lab settings may not be reliable indicators of the effects observed in the real world (Jerit, Barabas, and Clifford 2013). Moreover, the success of lab and survey experiments has policy implications limited to populations who are frequently online and use platforms such as Facebook and Mechanical Turk. This does not describe the vast majority of populations in developing countries,

who hold varying levels of digital literacy, are less likely to be avid Internet users, and more likely to fall prey to misinformation.

The body of academic research on measures to correct misinformation in the United States constitutes an important contribution to scholarship, but simultaneously highlights the dearth of high-quality field interventions and studies in developing contexts. The next sections outline the challenge posed by misinformation in developing countries and the need for solutions and interventions specific to those contexts.

3 Dissemination of Misinformation in India: The Supply

This study was conducted in May 2019 during the general election in India, the largest democratic exercise in the world. The 2019 contest was a reelection bid for Narendra Modi, leader of India's Hindu nationalist Bharatiya Janata Party (BJP). India is a parliamentary system but Narendra Modi's style of politics makes it akin to presidential elections with a high level of polarization, where not unlike Donald Trump, he "inspires either fervent loyalty or deep distrust" (Masih and Slater 2019).

This election was distinctive because it allowed for campaigning to be conducted over the Internet, and chat-based applications such as WhatsApp became a key communication tool for parties. For example, the BJP drew plans to have WhatsApp groups for each of India's 927,533 polling booths. A WhatsApp group can contain a maximum of 256 members, hence this communication strategy potentially reached 700 million voters. This, coupled with WhatsApp being the social media application of choice for over 90% of Internet users, led the BJP's social media chief to declare 2019 the year of India's first "WhatsApp elections" (Uttam 2018). Survey data from this period in India finds that one-sixth of respondents said they were members of a WhatsApp group chat started by a political leader or party (Kumar and Kumar 2018).

Unlike the United States where the focus has been on foreign-backed misinformation campaigns, political fake news circulating in India aimed at altering public opinion and elections appears to be largely domestically manufactured. The information spread on such political WhatsApp groups is not only partisan but also hate-filled and often false (Singh 2019). This trend is fueled by party workers themselves: ahead of the 2019 election, national parties hired armies of volunteers "whose job is to sit and forward messages" (Perrigo 2019). Singh (2019) reports that the BJP directed constituency-level volunteers to sort voters into groups created along religious and caste lines, even location, socioeconomic status and age, such that specific messages could be targeted to specific WhatsApp groups. Perrigo (2019) describes one such fake message on a pro-BJP group chat "Vote For Modi": a graphic image shows a man's body hanged by the neck outside a temple. "One more priest has been murdered," the accompanying message reads in Hindi. "Remember, the jihadis are not going to stop at just this." So entrenched is the political misinformation machinery in India that it resembles an industry where spreading false messages is incentivized. Then BJP President Amit Shah underscored these findings during a public ad-

dress in 2018: "We can keep making messages go viral, whether they are real or fake, sweet or sour" (Wire 2018). Thus misinformation is inherent political in India, and the creators of viral messages are often parties themselves.

4 Vulnerability to Misinformation in India: The Demand

WhatsApp group chats morph into havens for fake news in India. Four characteristics make such people in such group chats vulnerable to misinformation.

First, literacy and education rates are low across the developing world. India's literacy rate, along with its rate of formal education, is relatively low compared to other developing countries where misinformation has been shown to affect public opinion (Figure 1). Further, the sample site for this study – the state of Bihar in India – has historically had one of the lowest literacy rates within the country. Research has demonstrated a strong relationship between levels of education and vulnerability to misinformation. While people with higher levels of education have more accurate beliefs (Allcott and Gentzkow 2017), motivated reasoning gives them better tools to argue against counter-attitudinal information (Nyhan et al. 2019). We should thus expect a higher vulnerability on average to misinformation among populations with lower literacy and education.

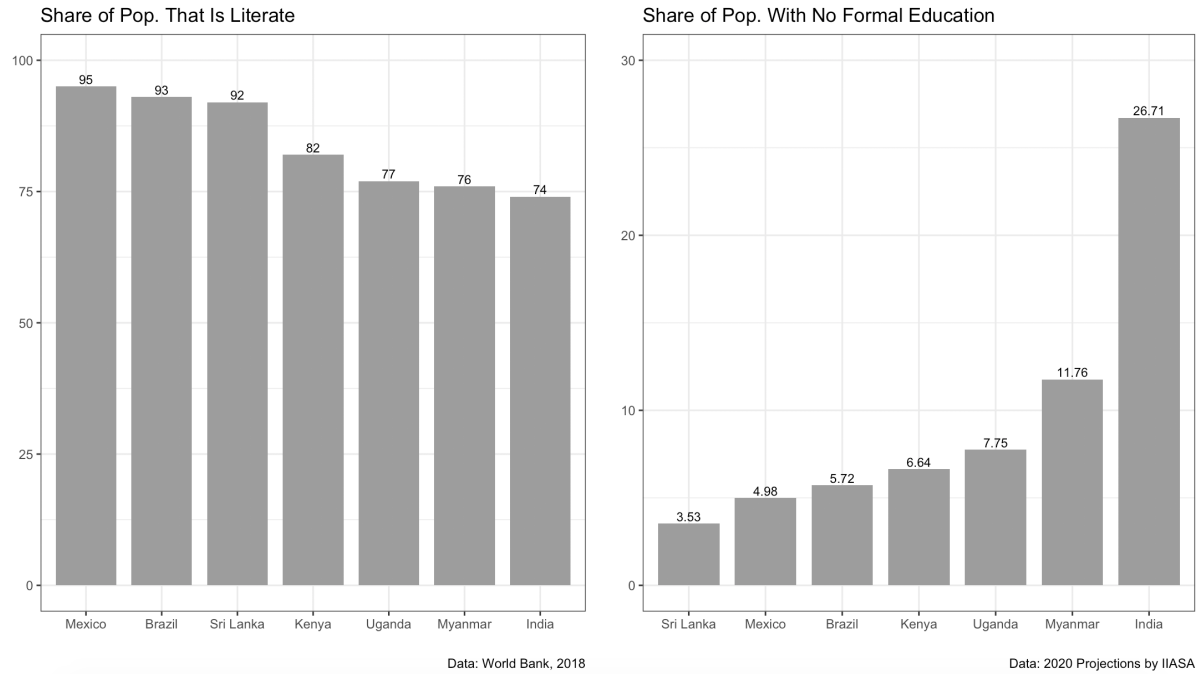

Figure 1: India Has Low Levels of Literacy and Education

Second, Internet access has exploded in the developing world. India, particularly, is digitizing faster than most mature and emerging economies. An average user in India currently consumes more than 8 GB of data per month, which exceeds the average in more technologically

advanced countries such as China and South Korea. This consumption is driven by the increasing availability and decreasing cost of high-speed connectivity and smartphones, and some of the world's cheapest data plans (Kaka et al. 2019). Internet penetration in India has increased exponentially over the past few years and Bihar – the sampling site for this study – saw an Internet connectivity growth of over 35% in 2018, the highest in the country (Mathur 2019).

Paradoxically, this leap in development coupled with the novelty and unfamiliarity with the Internet could make new users more vulnerable to information received online. 81% of users in India now own or have access to smartphones and most of these users report obtaining information and news through their phones (Devlin and Johnson 2019). But, research finds that obtaining accurate information on mobile devices is costlier than other mediums (Donner and Walton 2013) and that mobile-driven information attenuates attention paid to news (Dunaway et al. 2018). The example of Geeta highlights this aspect. Geeta lives in Arrah, Bihar in a small one-room home and recently bought a smartphone with Internet. I asked her if she thought information received over WhatsApp was factually accurate:

"This object [her Redmi phone] is only the size of my palm but is powerful enough to light up my home (...) Previously we would have to walk to the corner shop with a TV for the news. Now when this tiny device shines brightly and tells me what is happening in a city thousands of kilometers away, I feel like God is directly communicating with me" [translated from Hindi][1]

Geeta's example demonstrates that the novelty of digital media could increase vulnerability to all kinds of information. Survey data shows that countries like India have several "unconscious" users who are connected to the Internet without an awareness that they are going online (Silver and Smith 2019). Such users may be unaware of what the Internet is in a variety of ways. The expansion of Internet access and smartphone availability in India thus generate the illusion of a mythic nature of social media, underscoring a belief that if something is on the Internet, it must be true.

Third, online information in developing countries is disproportionately consumed on encrypted chat-based applications, such as WhatsApp. India is WhatsApp's biggest market in the world (with about 400 million users in mid-2019), but an important reason contributing to the app's popularity is also at the heart of the fake news problem: WhatsApp messages are private and protected by encryption. This means that no one, including the app developers and owners themselves, have access to see, read, filter, and analyze text messages. This feature prevents surveillance by design, such that tracing the source or the extent of spread of a message is close to impossible, making WhatsApp akin to a black hole of fake news. Critically, this means that interventions tested in the misinformation literature are inapplicable to WhatsApp – the platform cannot add a "disputed" tag to a dubious message, warning labels cannot appear beneath messages, suspect links cannot be removed. Top-down and platform-driven solutions are impractical in the case of private group chats on WhatsApp.

[1]Interview with Geeta, March 27, 2019. Unless noted otherwise, all individual names have been changed to protect the confidentiality of focus group participants.

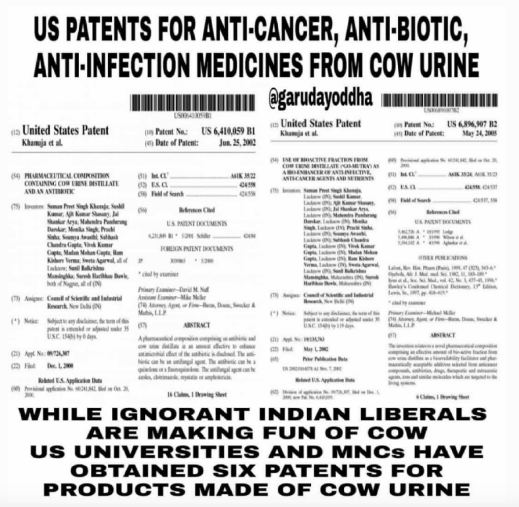

Figure 2: "Cow Urine Cures Cancer" Viral WhatsApp Rumor

Finally, the format of fake stories in India is mainly visual: much of what goes viral on WhatsApp constitutes photo-shopped images and manufactured videos. Misinformation in graphical and visual form is found to have increased salience, capable of retaining respondent attention to a higher degree (Flynn, Nyhan, and Reifler 2017). My intervention drew from a sampling of fake photoshopped images and pseudo-scientific narratives that became popular on WhatsApp in India in the months leading up to the election. Among these are false claims relating to the wondrous power of cows, along with rumors targeting minorities for storing beef or illegally slaughtering cows. Killing cows is sacrilege to many Hindus, illegal in some states, and is squarely a political and electoral issue in India (Ali 2020). According to Human Rights Watch, at least 44 people were killed in "cow-related violence" across 12 Indian states between May 2015 and December 2018. Most of them were Muslim. Figure 2 is an example of a fake story that circulated over WhatsApp prior to the election, claiming that cow urine cures cancer. On WhatsApp, fake news is almost never shared with a link – the image above was forwarded as is to thousands of users, making the original source unknown and impossible to debunk.

This image is also partisan in nature, highlighting differences between "Indian liberals", or those who do not support the right-wing BJP, and others on the political spectrum. Despite this, evidence on the power of partisanship and ideology as polarizing social identities in India is mixed. India's party system is not historically viewed as ideologically structured. Research finds that parties as not institutionalized (Chhibber, Jensenius, and Suryanarayan 2014), elections are highly volatile (Heath 2005), and the party system itself is not ideological (Ziegfeld 2016; Kitschelt and Wilkinson 2007; Chandra 2007). More recent literature, however, argues for the idea that Indians *are* reasonably well sorted ideologically into parties, and that while Indian politics has traditionally been viewed as clientelistic, it might be becoming more programmatic amongst certain groups (Chhibber and Verma 2018; Thachil 2014). Despite this, we know little about the origins of partisanship in India–whether it stems from transactional relationships

with parties, affect for leaders, ties to social groups, ideological leanings–or its stability. Given these findings, it is unclear whether theories of partisan motivated reasoning apply to the Indian context. However, misinformation in India is inherently political, and more so during elections when citizen attachments to political parties are heightened (Michelitch and Utych 2018). Putting these observations together, the strength of party identity as an indicator of attitudes in India is ambiguous, but timing of this study during the election, where party politics is arguably more polarizing, lends itself to interesting party effects on the treatment.

In sum, lower rates of education and literacy, new and accelerating Internet access, encrypted messaging applications, and visual forms of information make users in contexts such as India likely to fall for fake news.

5 Research Design

The nature of the supply of and vulnerability to misinformation in India demonstrates that preference-based accounts of misinformation are inadequate to understand the prevalence of fake news in developing countries, where citizens may not be fully aware about the existence of misinformation. Theoretical viewpoints that support such accounts may conclude that consumers who fall for misinformation are not biased, but simply lazy (Pennycook and Rand 2019). Such conclusions may mischaracterize media markets where consumers are not aware about the existence of fake news to begin with. Solutions to misinformation in developing contexts thus need to be updated to reflect the needs of a citizenry that is less educated, far newer to the Internet, and where primary modes of information consumption are encrypted group chat applications.

To address these challenges, I designed a pedagogical, in-person treatment with educative tools to address fake news in the Indian context. This setup is a natural extension of the existing solutions proposed by WhatsApp such as exhorting users to fact-check themselves, but allows for learning in a one-on-one session that provides users with tangible methods that they can use to ascertain the veracity of information.

5.1 Experimental Design

The intervention targeted to treatment group respondents was, by design, a bundled treatment incorporating several elements. It consisted of surveying a respondent in their home and undertaking the following activities in a 45-60 minute visit:

1. Introduction: The survey enumerator explained that they were there to have a conversation with the respondent about social media and politics, stated their affiliation to the survey organization, and explicitly stated that they were not affiliated with any political party.

2. Pre-treatment survey: Survey modules were administered to measure demographic and pre-treatment covariates including digital literacy, political knowledge, media trust, and

prior beliefs about misinformation.[2]

3. Pedagogical intervention: In this key part of the treatment, respondents learnt two concrete tools to identify misinformation.

   Performing reverse image searches: A large part of misinformation in India comprises of fake photos and videos, often drawn from one context and used to spread misinformation about another context or time. Reverse searching such images is an easy way identify their origins. For example, before the 2019 election, a viral image of a missile that claimed to kill over 300 terrorists in Pakistan circulated over WhatsApp chats in India. Reverse searching the image revealed it was a screenshot from a video game. As one focus group discussion conducted before the experiment revealed: *"the time stamp on the photo helped me realize that it is not current news; if this image has existed since 2009, it cannot be about the 2019 election"*.[3] Respondents can see the original source and time stamp on an image once it is fed back into Google, making this technique a uniquely useful and compelling tool given the nature of visual misinformation in India. Enumerators demonstrated two examples of this to respondents.

   Navigating a fact-checking site: Focus group discussions also revealed that while a minority of those surveyed knew about the existence of fact-checking websites in India, even fewer were able to name one. The second concrete tool in this intervention involved demonstrating to respondents how to navigate a fact-checking website, www.altnews.in, a non-profit fact-checking service in India. Enumerators pulled up the website's main page in Hindi,[4] explained the layout of the site, showed respondents where to find fact-checked viral fake stories, and demonstrated how to use the search bar.

4. Corrections and tips flyer: Enumerators next helped respondents apply these tools to fact-check four fake stories. Do to so enumerators displayed a flyer to respondents, the front side of which had descriptions of four recent viral political false stories. For each story, enumerators systematically corrected the fake news, explaining in each case why the story was untrue, what the correct version was, and what tools were used to determine veracity. The back side of the flyer contained six tips to reduce the spread of fake news. The enumerator read and explained each tip to respondents, gave them a copy of the flyer and exhorted them to make use of it.

   These tools were demonstrated to treatment group respondents only. Control group respondents were shown a placebo demonstration about plastic pollution, and were given a flyer containing tips to reduce plastic usage.

5. Comprehension Check: Enumerators lastly administered a comprehension check to measure whether the treatment was effective in the short-term.

[2] Summary statistics for all key variables are included in Table A.1
[3] Interview with Bharat, March 31, 2019.
[4] https://www.altnews.in/hindi/

Since encyrpted platforms put the onus of fact-checking on users, the intervention teaches respondents skills to do so themselves. This makes bottom-up intervention such as this one easily scalable – the treatment can spread to different platforms without having tech companies do the heavy lifting and can thus be implemented in a variety of contexts and settings.

For this study, respondents were randomized into one of three groups, two treatment and one placebo control. Table 1 summarizes the three groups.

Table 1: Experimental Treatments

Intervention	**Goal**
T1: Pedagogical Intervention + Pro-BJP flyer	Learning + corrections to 4 pro-BJP fake stories
T2: Pedagogical Intervention + Anti-BJP flyer	Learning + corrections to 4 anti-BJP fake stories
Control: Plastic Pollution Intervention + flyer	Learning + tips on plastic pollution

Respondents in both treatment groups received the pedagogical intervention. However, one group received corrections to four pro-BJP fake stories, the other received corrections to four anti-BJP fake stories. Besides differences in the stories that were fact-checked, the tips on the flyer remained the same for both treatment groups. Respondents in the placebo control group received a symmetric treatment where enumerators spoke about plastic pollution; they were left with a flyer on tips to reduce plastic usage. The fake news statements included in the treatment group flyers were drawn from a pool of stories fact checked for accuracy by altnews.in and boomlive.in. The partisan slant of each story was determined by a Mechanical Turk pre-test. To ensure balance across both treatment groups, stories with similar salience and subject matter were picked. The back of treatment flyers contained the same tips on how to verify information and spot fake news. The entire intervention was administered in Hindi. Figures C.1, C.2 and C.3 present the English-translated version of flyers distributed to respondents.

To control for potential imbalance in the sample, a randomized block design was used. The sample was divided into two subgroups based on partisan identity such that the variability within each subgroup was less than the variability between them. In the baseline survey, respondents were asked which national political party in India they felt closest to. Those respondents who chose the BJP were one block, those who chose anything but the BJP were another block. Within each block, respondents were randomly assigned to one of the three experimental groups described in Table 1. This design ensured that each treatment condition had an equal proportion of BJP and non-BJP partisans. Overall, the sample was equally divided between the two treatment and placebo control groups (i.e. one third of the sample in each of the three groups).

5.2 Sample and Timeline

The sample was drawn from the city of Gaya and its suburbs in the state of Bihar in India. Bihar has both the lowest literacy rate in the country as well as the highest rural penetration of mobile phones, making it a strong test-case for the intervention.

To maximize WhatsApp users who were familiar with the Internet in the sample, I restricted the sample based on 3 variables: respondents were required to own their own cellphone (i.e. not a shared household phone), have working Internet for 6 months prior to the survey, and WhatsApp was required to be downloaded on the phone. The final sample comprised of 1,224 respondents. Trained enumerators administered the intervention in a household visit rolled out in May 2019. Approximately two weeks after the intervention, the same respondents were revisited to conduct an endline survey and measure the outcomes of interest. Critically, respondents voted in the election between the two enumerator visits. Figure 3 summarizes the timeline for this study.

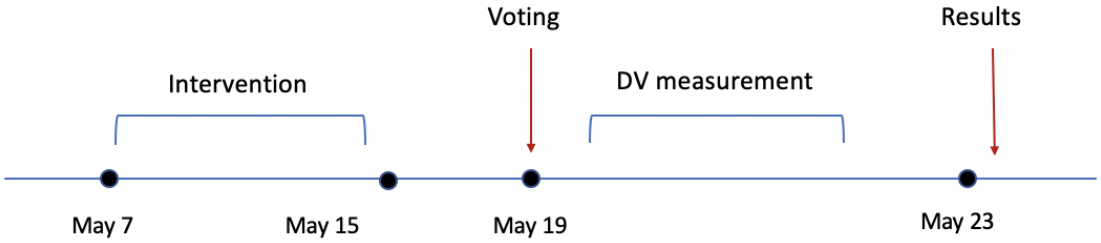

Figure 3: Experimental Timeline (May 2019)

The study took multiple steps in survey design and implementation to minimize exogenous shocks from election results. The timeline ensured that though respondents voted in the general election *after* the intervention, making voter turnout post-treatment, the endline survey to measure outcomes was conducted *before* election votes were counted and results were announced.[5] This timeline had the double advantage of ensuring that outcome measures was not impacted by the exogenous shock of results while also making sure respondents received the intervention before they voted, when political misinformation is arguably at its peak. At the end of the baseline survey, enumerators collected addresses and mobile numbers of respondents for subsequent rounds of the study and then immediately separated this contact information from the main body of the survey to maintain respondent privacy. Houses and respondents for the study were selected through a random walk procedure.[6]

[5] In India voting is staggered by constituency but ballots are counted after every constituency in the country has voted.

[6] Details about the sampling process are available in online appendix B.

5.3 Dependent Variables

In the endline survey, enumerators revisited the same respondents after they had voted in the 2019 elections. The same set of enumerators administered the intervention and the endline survey. However, enumerators were given a random set of household addresses for the endline survey so as to minimize the possibility of the same enumerator systematically interviewing the same respondent twice. Further, addresses and contact information were separated immediately from baseline survey data to ensure that enumerators only had contact information about respondents. During the baseline survey, 1306 respondents were administered the intervention. The enumerators successfully located 1224 of these respondents, resulting in an attrition rate of 6%. Importantly, nobody who was administered the intervention refused to answer the endline survey; the attrited group comprised only of respondents who enumerators were unable to contact at home after three tries.

The key outcome of interest is whether the intervention positively affected respondents' ability to identify fake news. To this end, respondents were shown a series of fourteen news stories.[7] These stories varied in content, salience, and critically, partisan slant. Half of the stories were pro-BJP in nature and the other half anti-BJP. [8] Each respondent saw all the fourteen stories, but the order in which they were shown was randomized.[9] Following each story, two primary dependent variables were measured:

1. Perceived accuracy of fake news identification, with the question "Do you believe this news story is fake?" (binary response, 1 if yes, 0 otherwise)

2. Confidence in identification of the story as fake or real, with the question "How confident are you that the story is real / fake?" (4-point scale, 1 = very confident, 4 = not confident at all)

A list of the fourteen stories shown to respondents is presented in Table D.1.[10]

5.4 Hypotheses and Estimation

This study aims to test whether teaching respondents information processing skills improves actual information processing. Hence, I hypothesize there will be a positive effect of the intervention for respondents assigned to any arm of the treatment group relative to placebo control.

[7] 12 were false and 2 were true. Given the countless, diverse array of stories that went viral in India during this time with perilous consequences, I chose to maximize on reducing belief in as many false stories as possible. Hence respondents were shown more false stories as part of the outcome measure (rather than a 50-50 split between true and false stories). Two true stories (each of different partisan slant) were included in the measure, and respondents were told that some of the stories were false and some true.

[8] Partisan slant of the news stories was determined with a Mechanical Turk pre-test.

[9] For field safety reasons, the endline survey was conducted offline and hence the order of appearance of the dependent variable stories was limited to 3 pre-determined random orders. A given enumerator had access to only one of the 3 random orders. As a robustness check, I replicate the main analysis with enumerator fixed effects. Results are presented in Tables E.1 and E.2.

[10] Online appendix D describes secondary dependent variables measured.

I also hypothesize that the individual effect of being assigned to each treatment will have a positive effect relative to placebo control:

Hypothesis 1: *Exposure to the media literacy intervention will increase ability to identify fake news relative to control.*
Hypothesis 2a: *Exposure to media literacy and pro-BJP corrections will increase ability to identify fake news.*
Hypothesis 2b: *Exposure to media literacy and anti-BJP corrections will increase ability to identify fake news.*

I estimate the following equations to test the main effect of the intervention:

$$FakeNewsId_i = \alpha + \beta_1 Intervention_i + \epsilon_i \tag{5.1}$$

$$FakeNewsId_i = \alpha + \beta_1 InterventionPro\text{-}BJP_i + \beta_2 InterventionAnti\text{-}BJP_i + \epsilon_i \tag{5.2}$$

In the equations, i represents the respondent, the *Intervention* variable in Equation 5.1 represents pooled assignment to media literacy intervention (relative to control). In Equation 5.2, the dependent variable is regressed on separate indicators for having received the intervention and pro-BJP corrections, or intervention and anti-BJP corrections, with the control condition as the omitted category. The dependent variable *FakeNewsId* counts the number of stories correctly identified as fake. *FakeNewsId* has been coded such that a positive estimated β_1 indicates an increase in the ability to identify fake news.

Beyond the average treatment effect, I expect treatment effects to differ conditional on a single factor previously identified in the literature as a significant predictor of information consumption: partisan identity. Given that the burgeoning literature on motivated reasoning posits that the most common sources of directional motivated reasoning are partisanship and strength of party identity (Nyhan and Reifler 2010), I expect that the treatment effect will be larger for politically incongruent information as compared to politically congruent information, relative to the control condition. A politically congruent condition manifests when corrections are pro-attitudinal, i.e, BJP partisans receiving corrections to anti-BJP fake stories, or non-BJP partisans receiving corrections to pro-BJP fake stories.

Hypothesis 3: *Effectiveness of the intervention will be higher for politically incongruent information compared to politically congruent information, relative to the control condition.*

To determine whether partisan identity moderates treatment effects, I test Hypothesis 3 with the following model:

$$FakeNewsId_i = \alpha + \beta_1 Intervention_i + \beta_2 Intervention_i * PartyID_i + \beta_3 PartyID_i + \epsilon_i \tag{5.3}$$

In Equation 5.3, *PartyID* is an indicator variable that takes on the value of 1 if the re-

spondent self-identified as a BJP supporter. The choice to code party identity as dichotomous was based on the nature of misinformation in India where false stories are perceived as either favoring or not favoring the BJP. A positive coefficient estimate for β_2 indicates an increase in the ability to identify fake stories among BJP partisans due to the treatment.

6 Data and Results

This section begins with descriptive analyses that demonstrate the extent of belief in fake news as well as partisan polarization in this belief.

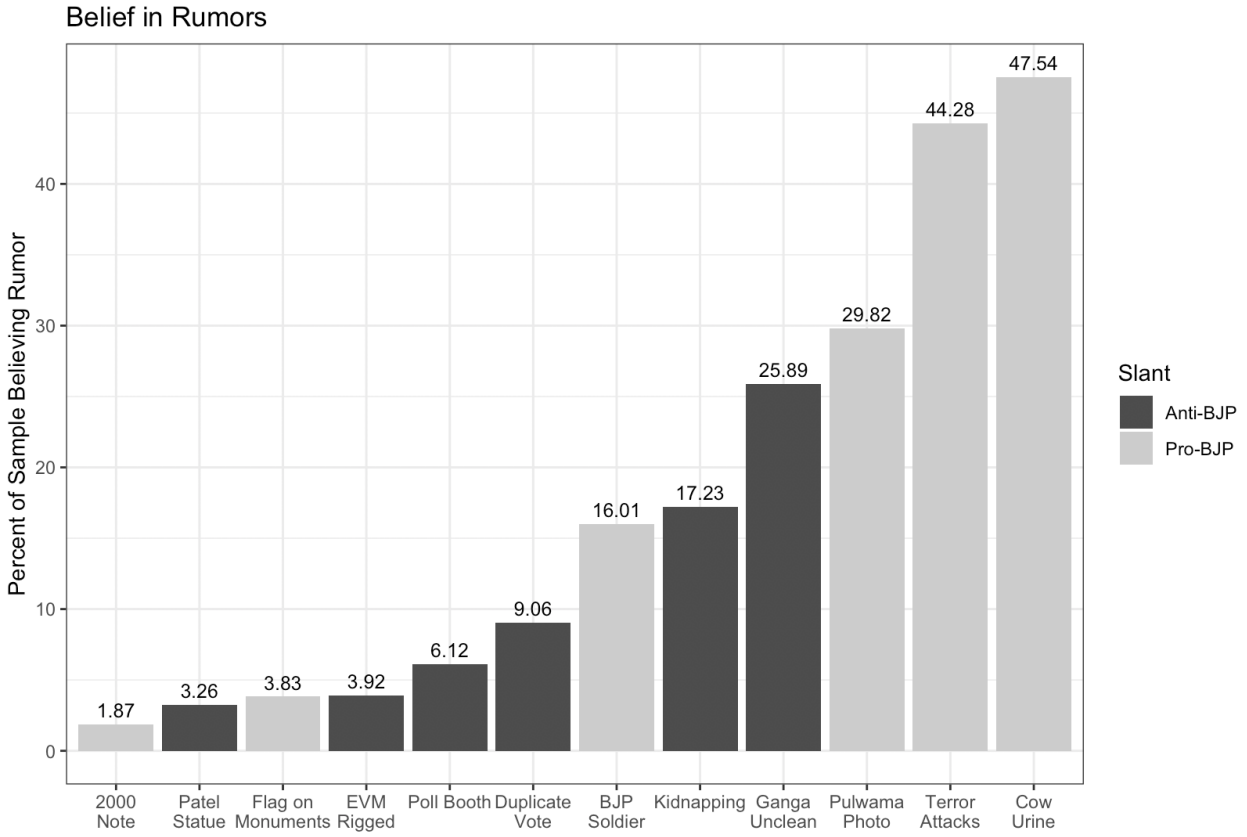

Figure 4: Percent of Sample Who Believe Rumors

Figure 4 lists the 12 false rumors used in the dependent variable measure in this study. This figure plots the share of respondents in the sample who believed each story to be true. Three aspects of the figure are striking. First, general belief in misinformation is low. For half of the 12 false stories, less than 10% of the sample thought they were true. Second, belief in pro-BJP fake news appears to be stronger, possibly alluding to its increased salience (Jerit and Barabas 2012), frequency of appearance on social media (Sinha, Sheikh, and Sidharth 2019), or to the presence of a higher proportion of BJP supporters in the sample. Third, belief in Pakistan

and foreign terrorism-related fake news (Pulwama Photo, Terror Attacks, BJP Soldier) is higher as compared to domestic issues, and belief in election tampering rumors is low overall (EVM Rigged, Poll Booth, Duplicate Vote). Lastly, respondents demonstrated the highest belief (47.54%) in the "Gomutra" rumor, the story that cow urine can cure cancer. Overall, across the 12 rumors, respondents correctly classified an average of 9.91 rumors.

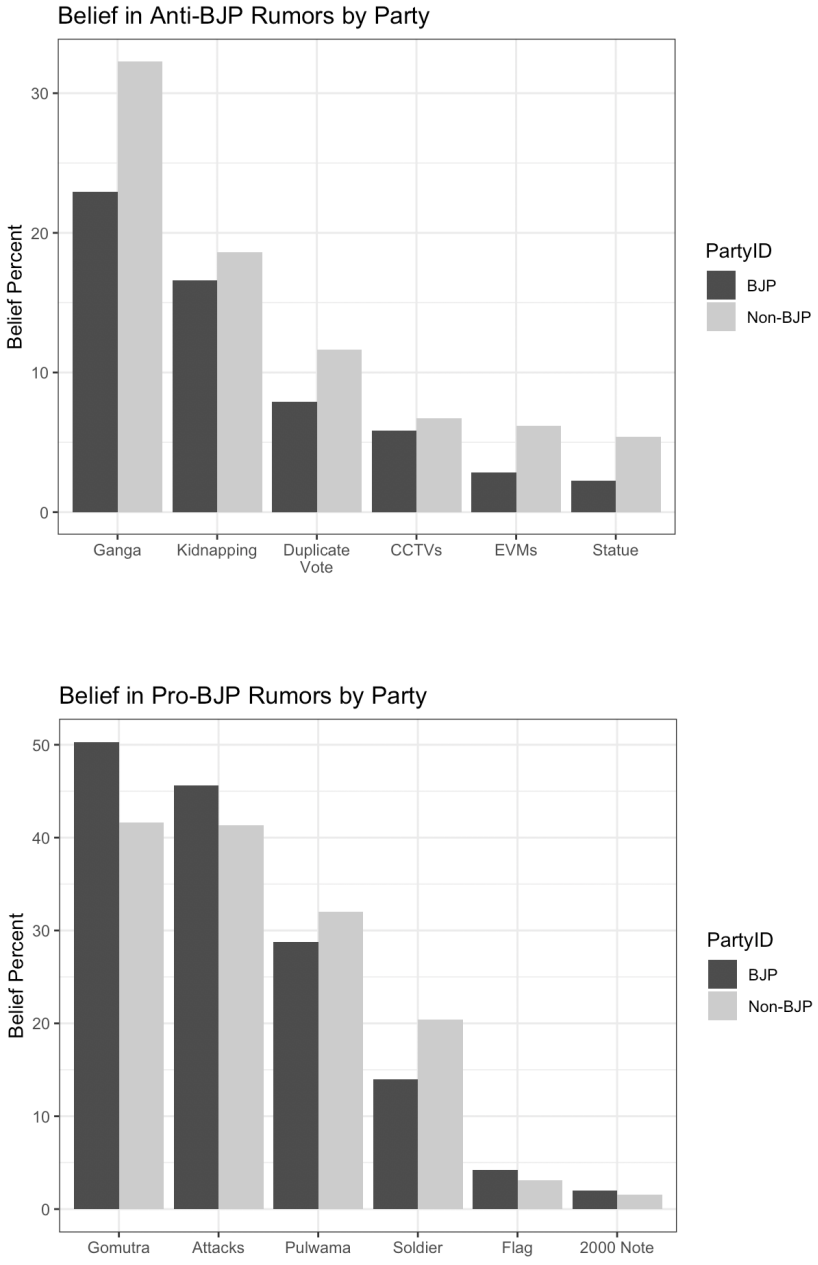

Figure 5: Belief in Rumors by Respondents' Party ID

Figure 5 plots respondent belief in rumors by self-reported partisan identity. For 10 out of the 12 partisan rumors, we see a correspondence between respondent party identity and pre-

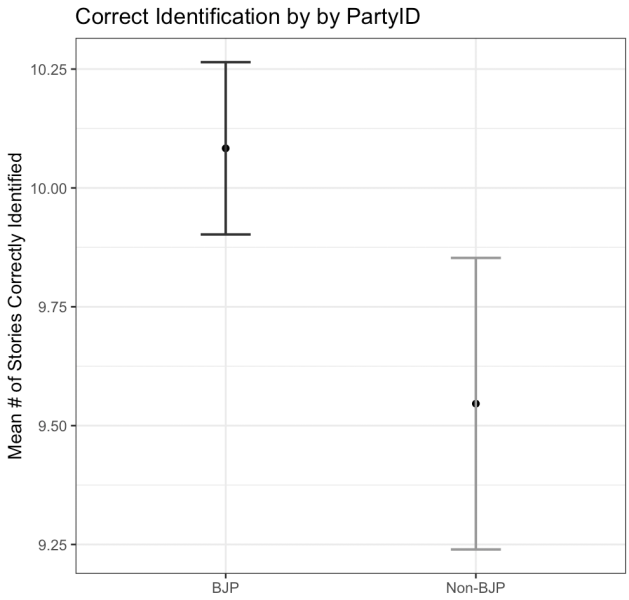

Figure 6: Misinformation Identification by Party ID
(Control Group Respondents)

tested political slant of the rumor. Though there is partisan sorting on belief in political rumors, the gap between BJP and non-BJP partisans in their beliefs is not as large as in the American case: the biggest gap appears in the case of the Unclean Ganga river rumor, where non-BJP partisans showed about 9 percentage points more belief in the rumor relative to BJP supporters. In contrast, Jardina and Traugott (2019) demonstrate that differences between Democrats and Republicans in their belief of the Obama birther rumor can be as large as 80 percentage points.

To identify differences between sub-populations in vulnerability to misinformation, I analyze the correlates of misinformation among control group respondents (N=406). To do this, I calculate the mean number of stories that were correctly identified, such that the average represents the baseline rate of identification ability in the absence of the intervention. The Y-axis is the mean number of stories correctly identified as false amongst the 12 fake stories presented to respondents.

In Figure 6 I plot differences in identification of fake stories by party. Respondents in the control group who self-identified as BJP supporters were significantly better at identifying fake stories than respondents who did not identify as BJP supporters. This observational result is striking – on the one hand, pro-BJP rumors are more likely to be believed by respondents, in line with descriptions of a right-wing advantage in producing misinformation (supply side). However, demand side results demonstrate that BJP supporters are better at identifying fake news. This finding bodes with observations that incentives to spread partisan misinformation has led parties like the BJP to form "cyber-armies" to disseminate information. Thus, while it is possible that BJP respondents are more aware of party-driven supply of misinformation, thereby

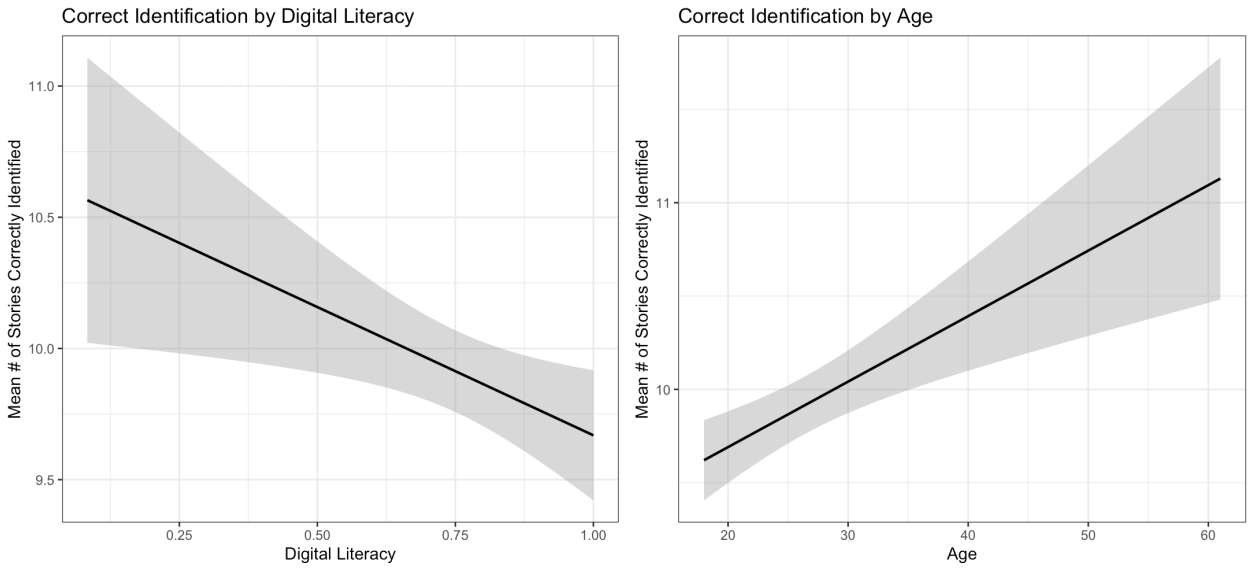

Figure 7: Misinformation Identification by Age and Digital Literacy
(Control Group Respondents)

being able to identify rumors at greater rates, their partisanship also makes them expressively believe pro-attitudinal rumors. These observational findings suggest the presence of partisan motivated reasoning in the Indian context.

In Figure 7 I plot the mean number of correctly identified stories by age and digital literacy. A consistent finding from political science research on misinformation is that older adults are most likely to engage with fake news sources (Grinberg et al. 2019). My findings demonstrate the opposite: younger adults are less likely to identify false stories, with the rate of identification growing with age. However, this sample is predominantly young: sampling conditions resulted in an uneven age distribution, with about 35% of respondents below age 22 and only about 6% of the sample above age 45.[11] Thus, the smaller sample of respondents above 50 might explain why the Indian context does not produce a similar result to the American context. While results from the U.S. context show a monotonically increasing age function with respect to vulnerability to misinformation (Barberá et al. 2015), this is contrary to what I find.

With respect to digital literacy, we observe in this data that increases in digital literacy are associated with lower identification of fake news. Literature in the American context shows that people who are less digitally literate are more likely to fall for fake news and clickbait (Munger et al. 2018). Hence similar to age, the Indian context presents a contradictory finding, made even more striking given the prevalence of misinformation. I measure digital literacy with eight (self-reported) ratings of degree of understanding of WhatsApp-related items, adapted for the Indian context from Hargittai (2005). While this measure demonstrates familiarity with the medium,

[11]The sampling conditions for this study required personal cellphone ownership with working WhatsApp and Internet. For more details on the sampling strategy and conditions see online appendix B.

decreases in identification suggest that such familiarity may not encompass cognitive tools necessary to discern fake news from real news. Arguably, those with higher digital literacy are more adept at sharing information while those with lower digital literacy may not have the skills to mass forward WhatsApp text messages to hundreds of contacts. Related to this, I also find that younger adults in my sample have significantly higher levels of digital literacy, despite being worse at identifying fake news.[12] However, familiarity with the way chat applications operate might lead respondents to feel less overwhelmed by the high-velocity online news environment, thereby sharing news at higher rates but not necessarily taking active means to differentiate between real and fake news (Joseph and Wihbey 2019). Taken together these results indicate that misinformation can spread despite increases in digital literacy, suggesting the need for educative interventions that teach people to engage better with the news.[13]

I now move to discussing experimental results. Enumerators administered a comprehension check at the end of the intervention to measure whether the treatment was effective in the short-term. Respondents were shown two false stories that were debunked by enumerators in the same house visit (as a part of the flyer with corrections). For each story, immediately after the treatment, respondents were asked to identify whether it was fake or not. Less than 5% of the sample for both stories incorrectly identified them as true, demonstrating that in the short run, respondents were able to successfully identify stories as fake after they had been debunked.

I estimate effects of the treatment on outcomes in a between-subjects design. All estimates are ordinary least square (OLS) regressions and empirical models are specified relying on random treatment assignment to control for potential confounders. First, I analyze data for the main effect of the intervention on ability to identify fake news. While research predicts that in-person and field interventions on media effects are likely to have stronger effects (Jerit, Barabas, and Clifford 2013; Flynn, Nyhan, and Reifler 2017), my findings from misinformation-prone India are less encouraging. Even with an in-person intervention, where enumerators spend close to one hour with each respondents to debunk and discuss misinformation and where respondents understood the intervention, I do not see significant increases in the ability to identify fake news as a function of teaching respondents media literacy tools.

Results are shown in Table 2. The key dependent variable in my analysis counts the number of stories that a respondent correctly identified as fake.[14] Columns 1 and 3 include stories that were classified from the pre-test as having a pro-BJP slant, Columns 2 and 4 include stories that were classified as having an anti-BJP slant. In order to estimate the pooled effect of the intervention, I construct a variable that takes on the value of 1 if a respondent received any literacy and fact-checking treatment (relative to 0 if the respondent was in the placebo control group). This effect of this pooled treatment is estimated in models (1) and (2). In models (3)

[12] See online appendix H for further analyses on this.

[13] Table G.1 presents regression results from including key pre-treatment covariates and demographic variables as controls.

[14] The dependent variable in these models counts the number of stories identified as fake out of a total of 12 false stories. I replicate these analysis where the dependent variable is the share of correctly identified stories given all fourteen stories, true and false, and find that the results hold. Analyses are in Tables F.1 and F.2.

and (4), I split the treatment into the pro-BJP corrections and the anti-BJP corrections (note both treatment conditions receive the same literacy intervention).

Table 2: Effect of Treatment on Ability to Identify Fake News

	Dependent variable: Number of Stories Identified as Fake			
	Pro-BJP Stories	Anti-BJP Stories	Pro-BJP Stories	Anti-BJP Stories
	(1)	(2)	(3)	(4)
Literacy Intervention	−0.004	0.004		
	(0.067)	(0.056)		
Literacy + Pro-BJP Fact-Check			0.013	0.013
			(0.078)	(0.065)
Literacy + Anti-BJP Fact-Check			−0.021	−0.006
			(0.078)	(0.065)
Constant	4.569***	5.342***	4.569***	5.342***
	(0.055)	(0.046)	(0.055)	(0.046)
Observations	1,224	1,224	1,224	1,224
R^2	0.00000	0.00000	0.0002	0.0001
Adjusted R^2	−0.001	−0.001	−0.001	−0.002
Res. Std. Error	1.110 (df = 1222)	0.925 (df = 1222)	1.111 (df = 1221)	0.926 (df = 1221)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2 demonstrates that the intervention did not increase fake news identification ability on average. Splitting the treatment into its component parts (each compared to placebo control) yields similar results. I find no evidence that an hour-long pedagogical intervention increased ability to identify fake news among respondents in Bihar, India. It appears that the intervention was not successful at explaining the dependent variable as there is no evidence that it had desirable effects on discerning the veracity of news stories after a period of time. The ability to update one's priors in response to factual information is privately and socially valuable, and hence the fact that a strong, in-person treatment does not change opinions demonstrates the resilience of fake news in India. Priors about fake news in this context appear resistant to change but, as I demonstrate below, this does not preclude moderating effects of partisan identity.

This null result could manifest for a number of reasons. The context of this experiment was arguably the strongest test case for the intervention: it took place in Bihar (state with the lowest literacy rate in the entire country), and during the election (when the salience of partisan misinformation is higher). The timing of the intervention during the election makes it possible that respondents had already seen many of these fake stories and internalized a response. It is also possible that citizen attachments to parties were heightened during this time, such that beliefs about information were a function of ethnicity, religion, and partisanship. Finally, the generally low levels of belief in rumors (as in Figure 4) suggest that respondents' attitudes were already near their ceiling, making it potentially harder for the intervention to succeed.

I now turn to the analysis of heterogeneous effects of partisan identity. Table 3 presents results. In Column 1 I estimate the effect of receiving the treatment for BJP supporters on ability to identify pro-BJP fake stories, Column 2 does the same with anti-BJP fake stories. The treatment variable for both models pools across receiving any treatment relative to control.

Table 3: Effect of Treatment x Party on Ability to Identify Fake News

	Dependent variable: Number of Stories Identified as Fake	
	Pro-BJP Stories	Anti-BJP Stories
	(1)	(2)
Literacy Intervention	0.277**	0.091
	(0.119)	(0.099)
BJP Supporter	0.226*	0.311***
	(0.118)	(0.098)
Literacy Intervention x BJP Supporter	−0.412***	−0.130
	(0.144)	(0.120)
Constant	4.415***	5.131***
	(0.097)	(0.081)
Observations	1,224	1,224
R^2	0.007	0.014
Adjusted R^2	0.005	0.011
Residual Std. Error (df = 1220)	1.107	0.920
F Statistic (df = 3; 1220)	2.892**	5.651***
Note:		*p<0.1; **p<0.05; ***p<0.01

Results are striking: while there was no average treatment affect, the interaction effect of the treatment on BJP partisans produces a negative effect on the ability to identify fake news. This effect is seen for pro-BJP stories. From Column 1 of Table 3 it is clear that the main effect of being a BJP supporter is positive, that is, the coefficient of 0.226 is the effect of being a BJP supporter in the absence of the intervention. Thus within the control group, BJP partisans are able to successfully identify 0.226 additional fake stories, relative to non-BJP partisans. We also see that the main effect of the intervention is positive and significant. This means that for non-BJP partisans, receiving the intervention led to an identification of 0.277 additional pro-BJP stories as fake.

The interaction between receiving the intervention and party identity highlights the interactive conditional effect of these variables. Visualizing this effect in Figure 8, where I graph the predicted values from the interaction model in Equation 5.3, it is evident that the treatment had contradictory effects conditional on party identity (for the set of pro-BJP stories). The intercept for BJP partisans is higher, demonstrating better identification skills ex-ante, in the absence of the treatment. However, treatment group respondents who identify as BJP partisans show a significant decrease in their ability to identify fake stories, while treatment group respondents

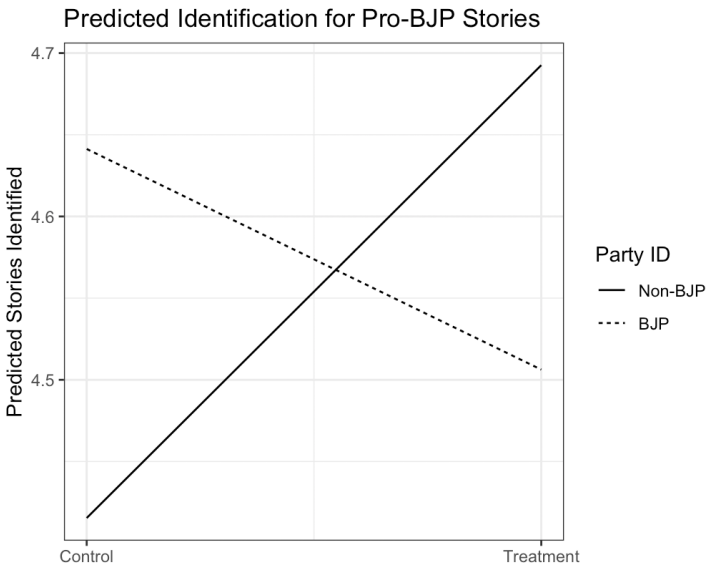

Figure 8: Predicted Identification of Pro-BJP Stories

who do not identify as BJP partisans show an increase in their ability to identify fake stories. Thus the treatment was successful with non-BJP partisans, and backfired for BJP partisans. Importantly, these effects obtain only for the set of fake stories that is pro-BJP in slant (implying that their corrections could be perceived as pro-attitudinal for non-BJP partisans). In Figure 9 I graph the interaction for the set of dependent variable stories that are anti-BJP in slant. While the relationships in this graph are directionally similar, they are smaller in magnitude and not significant. Importantly, fact-checking is much more effective for anti-BJP stories than for pro-BJP stories (note that the effects are much larger). Pro-BJP stories are more likely to be identified as fake in the control, but the treatment is weaker for this subset of stories. Taken together, these results imply that non-BJP respondents were able to successfully apply the treatment to identify pro-attitudinal corrections. But for BJP partisans, given that these corrections are not consistent with their partisan identity, the treatment backfires.

Despite the negative relationship between digital literacy and successful identification in the observational data, heterogeneous effects of the treatment demonstrate that certain subpopulations in the sample could successfully learn from the intervention and improve information processing, suggesting that cognitive detection of real from fake news operates orthogonally to digital literacy. However, finding that higher levels of identification (in the control group for BJP respondents) were made *worse* as a function of the treatment demonstrates the existence of partisan motivated reasoning in the Indian context. While it seems rational for strong partisans to deliberately discredit the out-party whilst simultaneously accepting misinformation that paints their party in a good light, it is worth investigating why this effect manifests as a function of the treatment. One explanation for these results is that in asking respondents questions about particular stories, those stories are made salient in their minds, likely to a greater extent for stories that were salient in the news to begin with (pro-BJP fake news) and for stronger partisans (BJP

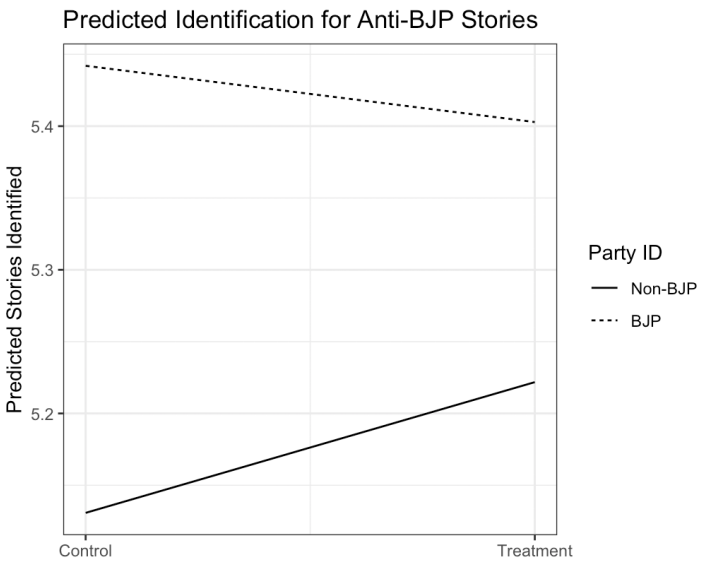

Figure 9: Predicted Identification of Pro-BJP Stories

supporters). Thus the negative coefficient on Treatment × BJP could be an exaggerated response due to partisans' stronger tendency to want to "cheerlead" for their party, as opposed to control group BJP respondents (Gerber and Huber 2009; Prior, Sood, and Khanna 2015). Related to this argument is evidence on the effects of motivated numeracy. Kahan et al. (2017) demonstrate that respondents highest in numeracy – a measure of the ability and disposition to make use of quantitative information – are better at drawing inferences from data (about a neutral topic). However, their responses become politically polarized-and even less accurate–when the same data is presented as results from a study on a partisan topic. In this experiment, BJP respondents were not significantly different from non-BJP respondents in self-reported levels of digital literacy. However, the fact that politically polarizing information can reverse the effects of baseline misinformation identification skills demonstrates the strong presence of partisan motivated reasoning in this context. I examine this result further in the Discussion.

6.1 Discussion

The most striking finding to emerge from this study demonstrates that the intervention improved misinformation identification skills for one set of respondents (non-BJP respondents) but not another (BJP partisans), where it backfired. Paralleling results seen in developed contexts, the perceptual screen (Campbell et al. 1960) of BJP partisanship shaped how respondents interacted with this treatment. This finding supports results demonstrating that citizen attachments to political parties are heightened during elections (Michelitch and Utych 2018) and that strong partisans engage in strategic ignorance, pushing away information and facts that get in the way of feelings (McGoey 2012). Future work should examine treatments such as this one in non-electoral contexts where the salience of partisanship may be lower, resulting in smaller differences be-

tween parties. Nevertheless, my findings suggest that even in democracies with weaker partisan identification, citizens still engage in motivated reasoning. This has important implications beyond the study of fact-checking and extends more broadly to how Indian citizens make political judgements.

Future research should also evaluate the effectiveness of bottom-up pedagogical interventions of various forms. These have the advantage of being more scalable as they do not depend on tech companies to do the heavy-lifting and can consequently be applied to a wide range of platforms. They also have the advantage of imparting cognitive skills to respondents in contexts where respondents are proficient at using their phone but may be "unconscious" Internet users. The descriptive findings from this study suggest that certain subgroups were able to learn from the intervention and improve identification skills. But, partisan differences in these beliefs (Figure 5) indicate that new information was unlikely to change attitudes of certain partisans, suggesting that bottom-up learning interventions are likely to work better on respondents who are yet to form strong political attitudes. Social scientists believe that the development of political identity during childhood appears to profoundly influence future political decision making (Campbell et al. 1960; Green, Palmquist, and Schickler 2004). Research demonstrates that children are most impressionable when they are adolescents and tend to form stable party identities in their teenage years (Jennings and Niemi 1968; Margolis 2018). Keeping this in mind, an apt example of one such target population for bottom-up pedagogical interventions is schoolchildren. In developing countries where people are more likely to use encrypted forms of social media, the most promising long-term solutions involve long-term pedagogical programs about the negative effects of misinformation and tools to combat it.

7 Conclusion

Misinformation campaigns have the capacity to affect opinions and elections across the world. Purveyors and victims of misinformation and hyper-partisan messaging are no more just individuals with low digital literacy skills, people who are uninformed, Internet scammers, or Russian trolls. A global rise in polarization has meant that the creators and contributors of fake news include party workers, stakeholders and politicians themselves. As the world moves to deal with the COVID-19 crisis, we are engulfed in a new deluge of misinformation in hyper-partisan and polarized environments, where traditionally non-political issues are also deeply politicized. The rise of polarization amidst a global pandemic underscores the need to identify robust strategies to counter the pernicious effects of misinformation, especially in societies where it is spread on encrypted platforms and where the stakes are as high as violence.

In this paper, I present new evidence on belief in popular fake news stories in India in the context of the 2019 general elections. Given the encryption and private nature of WhatsApp, the most popular social networking application in several developing countries, I design a pedagogical intervention to foster bottom-up skills training to identify misinformation. Using

tools specifically designed for the Indian context such as reverse image searches, I administer in-person skills training to 1224 respondents in Bihar, India in a field experiment. I find that this grassroots-level pedagogical intervention in has little effect on respondent ability to identify fake news on average. But, the partisanship and polarization of BJP supporters appears stickier than that of their out-partisans. Non-BJP supporters in the sample receive the treatment and apply it to identify fake news at a higher level, demonstrating that cognitive skills can be improved as a function of the treatment. However, for BJP partisans, receiving the treatment leads to a significant decrease in identification ability, especially for pro-attitudinal stories. Thus the intervention worked in opposing ways conditional on party identity: non-BJP respondents were able to successfully learn from the treatment, but the presence of partisan motivated reasoning subdued any positive effects for BJP respondents.

The existence of motivated reasoning is a surprising result in a country with traditionally weak party ties and non-ideological party systems. Democratic citizens have a stake in dispelling rumors and falsehoods, but in societies with polarized social groups, individuals also have a stake in maintaining their personal standing in social groups that matter to them (Kahan et al. 2017). The finding that the intervention worked on a subset of respondents underscores the fact that the training was not strong enough to overcome the effects of group identity for BJP respondents. Theoretically, this result is similar to research that finds that identity protective cognition, a type of motivated reasoning, increases pressure to form group-congruent beliefs and steers individuals away from beliefs that could alienate them from others they are similar to (Sherman and Cohen 2006; Giner-Sorolla and Chaiken 1997). Practically, the result calls for a revision of findings on party identity in India, as it demonstrates the presence of motivated reasoning in electoral settings.

The effects of party identity in this setting are arguably observable because of the rise of hyper-partisan and polarizing parties in India and across the world. It underscores a broader phenomenon of populist parties and narratives, resulting in societies where information is weaponized to divide polarized voters. While elections are times when political discourse is polarized and partisanship salience is heightened, these findings stress the need for more systematic research into motivated reasoning and polarization in societies that have traditionally been non-ideological and where encrypted forms of social media take center stage in the spread of misinformation. These encrypted mediums necessitate that top-down interventions to fight misinformation are rendered ineffective by design. Thus, future research should test the effect of long-term learning and skills training to counter misinformation.

My findings are somber for the implications of misinformation on democracy. While there are many complex factors responsible for the rise of illiberal populist governments around the world, the rise in populism correlates with the explosion in access to the Internet, suggesting that an increase in information has not been healthy in contexts where more information equals more false information. Just as Gresham's Law posits that bad money drives out good, bad (false) information has the effect of invalidating the vast advances made in the spread of technology

and free flow of information. Fake news threatens democracy, and there has never been a more pressing time to critically evaluate the effectiveness of solutions to fight misinformation.

References

Ali, Mohammed. 2020. "The Rise of a Hindu Vigilante in the Age of WhatsApp and Modi." *Wired*. April 14, 2020. https://www.wired.com/story/indias-frightening-descent-social-media-terror/.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–36.

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* 26 (10): 1531–1542.

Bode, Leticia, and Emily K Vraga. 2015. "In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media." *Journal of Communication* 65 (4): 619–638.

Campbell, Angus, Philip E Converse, Warren E Miller, and Donald E Stokes. 1960. *The American Voter*. New York: John Wiley.

Chan, Man-pui Sally, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. "Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation." *Psychological Science* 28 (11): 1531–1546.

Chandra, Kanchan. 2007. *Why Ethnic Parties Succeed: Patronage and Ethnic Headcounts in India*. New York: Cambridge University Press.

Chhibber, Pradeep, Francesca Refsum Jensenius, and Pavithra Suryanarayan. 2014. "Party organization and party proliferation in India." *Party Politics* 20 (4): 489–505.

Chhibber, Pradeep, and Rahul Verma. 2018. *Ideology and Identity: The Changing Party Systems of India*. New York: Oxford University Press.

Clayton, Katherine, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan et al. 2019. "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media." *Political Behavior*: 1–23.

Devlin, Kat, and Courtney Johnson. 2019. "Indian elections nearing amid frustration with politics, concerns about misinformation." *Pew Research Center*. March 25, 2019. https://www.pewresearch.org/fact-tank/2019/03/25/indian-elections-nearing-amid-frustration-with-politics-concerns-about-misinformation/.

Donner, Jonathan, and Marion Walton. 2013. "Your phone has internet-why are you at a library PC? Re-imagining public access in the mobile internet era." In *IFIP Conference on Human-Computer Interaction*. Springer pp. 347–364. https://link.springer.com/content/pdf/10.1007/978-3-642-40483-2-25.pdf.

Dunaway, Johanna, Kathleen Searles, Mingxiao Sui, and Newly Paul. 2018. "News Attention in a Mobile Era." *Journal of Computer-Mediated Communication* 23 (2): 107–124.

Flynn, D.J. 2016. "The Scope and Correlates of Political Misperceptions in the Mass Public." *Working Paper*. http://djflynn.org/wp-content/uploads/2016/08/Flynn-APSA2016.pdf.

Flynn, D.J., Brendan Nyhan, and Jason Reifler. 2017. "The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics." *Political Psychology* 38: 127–150.

Freed, Gary L, Sarah J Clark, Amy T Butchart, Dianne C Singer, and Matthew M Davis. 2010. "Parental Vaccine Safety Concerns in 2009." *Pediatrics* 125 (4): 654–659.

Fridkin, Kim, Patrick J Kenney, and Amanda Wintersieck. 2015. "Liar, Liar, Pants on Fire: How Fact-Checking Influences Citizens' Reactions to Negative Advertising." *Political Communication* 32 (1): 127–151.

Gentzkow, Matthew, Jesse M Shapiro, and Daniel F Stone. 2015. "Media Bias in the Marketplace: Theory." In *Handbook of Media Economics*. Vol. 1. Elsevier.

Gerber, Alan S, and Gregory A Huber. 2009. "Partisanship and Economic Behavior: Do Partisan Differences in Economic Forecasts Predict Real Economic Behavior?" *American Political Science Review* 103 (3): 407–426.

Giner-Sorolla, Roger, and Sheily Chaiken. 1997. "Selective Use of Heuristic and Systematic Processing Under Fefense Motivation." *Personality and Social Psychology Bulletin* 23 (1): 84–97.

Green, Donald P, Bradley Palmquist, and Eric Schickler. 2004. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. New Haven: Yale University Press.

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

Hargittai, Eszter. 2005. "Survey Measures of Web-Oriented Digital Literacy." *Social Science Computer Review* 23 (3): 371–379.

Heath, Oliver. 2005. "Party systems, political cleavages and electoral volatility in India: A state-wise analysis, 1998–1999." *Electoral Studies* 24 (2): 177–199.

Hochschild, Jennifer L, and Katherine Levine Einstein. 2015. *Do Facts Matter?: Information and Misinformation in American Politics*. Norman, OK: University of Oklahoma Press.

Jardina, Ashley, and Michael Traugott. 2019. "The Genesis of the Birther Rumor: Partisanship, Racial Attitudes, and Political Knowledge." *Journal of Race, Ethnicity and Politics* 4 (1): 60–80.

Jennings, M Kent, and Richard G Niemi. 1968. "The Transmission of Political Values from Parent to Child." *American Political Science Review* 62 (1): 169–184.

Jerit, Jennifer, and Jason Barabas. 2012. "Partisan Perceptual Bias and the Information Environment." *The Journal of Politics* 74 (3): 672–684.

Jerit, Jennifer, Jason Barabas, and Scott Clifford. 2013. "Comparing Contemporaneous Laboratory and Field Experiments on Media Effects." *Public Opinion Quarterly* 77 (1): 256–282.

Joseph, Kenneth, and John Wihbey. 2019. "Breaking News and Younger Twitter Users: Comparing Self-Reported Motivations to Online Behavior." In *Proceedings of the 10th International Conference on Social Media and Society*. pp. 83–91. https://dl.acm.org/doi/pdf/10.1145/3328529.3328548.

Kahan, Dan M, Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. 2017. "Motivated Numeracy and Enlightened Self-Government." *Behavioural Public Policy* 1 (1): 54–86.

Kaka, Noshir, Anu Madgavkar, Alok Kshirsagar, Rajat Gupta, James Manyika, Kushe Bahl, and Shishir Gupta. 2019. "Digital India: Technology to transform a connected nation." *McKinsey Global Institute*. March, 2019. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/digital-india-technology-to-transform-a-connected-nation.

Kitschelt, Herbert, and Steven I Wilkinson. 2007. *Patrons, Clients and Policies: Patterns of Democratic Accountability and Political Competition*. Cambridge: Cambridge University Press.

Kumar, Sanjay, and Pranav Kumar. 2018. "How widespread is WhatsApp's usage in India?" *Live Mint*. July 18, 2018. https://www.livemint.com/Technology/O6DLmIibCCV5luEG9XuJWL/How-widespread-is-WhatsApps-usage-in-India.html.

Kunda, Ziva. 1990. "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (3): 480–498.

Margolis, Michele F. 2018. *From Politics to The Pews: How Partisanship and The Political Environment Shape Religious Identity*. Chicago: The University of Chicago Press.

Masih, Niha, and Joanna Slater. 2019. "U.S.-style polarization has arrived in India. Modi is at the heart of the divide." *The Washington Post*. May 20, 2019. https://www.washingtonpost.com/world/asiapacific/divided-families-and-tense-silences-us-style-polarization-arrives-in-india/2019/05/18/story.html.

Mathur, Nandita. 2019. "India's internet base crosses 500 million mark, driven by Rural India." *Live Mint*. March 11, 2019. https://www.livemint.com/industry/telecom/internet-users-exceed-500-million-rural-india-driving-growth-report-1552300847307.html.

McGoey, Linsey. 2012. "The logic of strategic ignorance." *The British Journal of Sociology* 63 (3): 553–576.

Michelitch, Kristin, and Stephen Utych. 2018. "Electoral Cycle Fluctuations in Partisanship: Global Evidence from Eighty-Six Countries." *The Journal of Politics* 80 (2): 412–427.

Mosseri, Adam. 2017. "A New Educational Tool Against Misinformation." *Facebook*. April 6, 2017. https://about.fb.com/news/2017/04/a-new-educational-tool-against-misinformation/.

Munger, Kevin, Mario Luca, Jonathan Nagler, and Joshua Tucker. 2018. "Everyone On Mechanical Turk is Above a Threshold of Digital Literacy: Sampling Strategies for Studying Digital Media Effects." *Working Paper*. https://csdp.princeton.edu/sites/csdp/files/media/munger-mturk-digital-literacy-note.pdf.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas Wood. 2019. "Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability." *Political Behavior*: 1–22.

Nyhan, Brendan, and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32 (2): 303–330.

Nyhan, Brendan, and Jason Reifler. 2012. "Misinformation and fact-checking: Research findings from Social Science." *New America Foundation Research Paper*. https://www.dartmouth.edu/nyhan/Misinformation-and-Fact-checking.pdf.

Pennycook, Gordon, and David G Rand. 2019. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition* 188: 39–50.

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): 1865–1880.

Perrigo, Billy. 2019. "How Volunteers for India's Ruling Party Are Using WhatsApp to Fuel Fake News Ahead of Elections." *TIME*. January 25, 2019. https://time.com/5512032/whatsapp-india-election-2019/.

Poonam, Snigdha, and Samarth Bansal. 2019. "Misinformation Is Endangering India's Election." *The Atlantic*. April 1, 2019. https://www.theatlantic.com/international/archive/2019/04/india-misinformation-election-fake-news/586123/.

Prior, Markus, Gaurav Sood, and Kabir Khanna. 2015. "You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions." *Quarterly Journal of Political Science* 10 (4): 489–518.

Sherman, David K, and Geoffrey L Cohen. 2006. "The Psychology of Self-Defense: Self-Affirmation Theory." *Advances in Experimental Social Psychology* 38: 183–242.

Silver, Laura, and Aaron Smith. 2019. "In some countries, many use the internet without realizing it." *Pew Research Center*. May 02, 2019. https://www.pewresearch.org/fact-tank/2019/05/02/in-some-countries-many-use-the-internet-without-realizing-it/.

Singh, Shivam Shankar. 2019. *How to win an Indian election: What political parties don't want you to know*. Gurgaon: Penguin Random House.

Sinha, Pratik, Sumaiya Sheikh, and Arjun Sidharth. 2019. *India Misinformed: The True Story*. Noida: HarperCollins India.

Smith, Jeff, Grace Jackson, and Seetha Raj. 2017. "Designing Against Misinformation." *Medium*. December 20, 2017. https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2.

Taber, Charles S, and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50 (3): 755–769.

Thachil, Tariq. 2014. "Elite Parties and Poor Voters: Theory and Evidence from India." *American Political Science Review* 108 (2): 454–477.

Uttam, Kumar. 2018. "For PM Modi's 2019 campaign, BJP readies its WhatsApp plan." *Hindustan Times*. September 29, 2018. https://www.hindustantimes.com/india-news/bjp-plans-a-whatsapp-campaign-for-2019-lok-sabha-election/story-lHQBYbxwXHaChc7Akk6hcI.html.

Wire. 2018. "Real or Fake, We Can Make Any Message Go Viral: Amit Shah to BJP Social Media Volunteers." *The Wire*. September 26, 2018. https://thewire.in/politics/amit-shah-bjp-fake-social-media-messages.

Ziegfeld, Adam. 2016. *Why Regional Parties?* New York: Cambridge University Press.

Online Appendices for
Educative Interventions to Combat Misinformation: Evidence From a Field Experiment in India

Contents

A Summary Statistics

Table A.1 provides summary statistics for key variables in this study. Literacy Intervention is a dummy variable indicating random assignment to both treatment groups relative to control. BJP supporter is a dummy variable indicating respondents' self-reported support for the BJP relative to all other parties. Accurate Priors measures prior beliefs in veracity of news with a battery of four stories (two true and two false); for each story respondents are asked to discern the veracity on a 3-point scale. The variable Accurate Priors calculates the mean accuracy rating across all four stories. Digital Literacy is measured through eight five-point (self-reported) ratings of degree of understanding of WhatsApp-related items. The variable Digital Literacy calculates the mean level of literacy across the eight items. Political Knowledge is measured by a battery of 6 questions of varying difficulty on local and national politics in India; the variable Political Knowledge counts the number of correct answers. WhatsApp Use Frequency measures how frequently respondents use WhatsApp on a 7-point scale ranging from a few times a month to a few times a day. Trust in WhatsApp measures respondents' level of trust in WhatsApp as an accurate medium of receiving news about politics, on a four-point scale.

Table A.1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Median	Max
Literacy Intervention	1,224	0.668	0.471	0	1	1
BJP Supporter	1,224	0.684	0.465	0	1	1
Accurate Priors	1,158	0.695	0.196	0	0.750	1
Digital Literacy	1,224	0.758	0.194	0.083	0.833	1
Political Knowledge	1,224	5.000	1.135	0	5	6
WhatsApp Use Frequency	1,224	6.068	0.952	1	6	7
Trust in WhatsApp	1,224	2.729	0.821	1	3	4
Education	1,224	9.388	2.652	1	9	13
Age	1,224	26.646	9.182	18	24	85
Male	1,224	0.911	0.285	0	1	1
Hindu	1,224	0.837	0.369	0	1	1

B Survey and Sampling Design

The primary sampling unit, the city of Gaya and its suburbs in Bihar, consists of several electoral polling booths (smallest administrative units). Out of the total number of polling booths, a random sample of 85 polling booths were selected (through a random number generator in the statistical framework R) to serve as enumeration areas.

Within each enumeration area, enumerators were instructed to survey 10-12 households following a random walk procedure. This methodology has the benefits of fast implementation and unpredictability of movement and was chosen over traditional listing methods so that enumerators could spend as little time in the field as possible given the potential for electoral violence. Surveying households within each chosen polling booth area involved choosing a starting point and then proceeding along a path, selecting every k^{th} household. I followed the method similar to that used by the Afrobarometer surveys of picking a sample starting point and then choosing a landmark as near as possible to the sample starting point. Landmarks could be street corners, schools, or water sources, and field enumerators were instructed to randomly rotate the choice of such landmarks. From the landmark starting point, the field enumerator walked in a designated direction away from the landmark and selected the tenth household for the survey, counting houses on both the left and the right. Once they left their first interview they continued in the same direction, selecting the next household after another interval of 10. If the settlement came to an end and there were no more houses, the field enumerator turned at right angles to the right and kept walking, continuing to count until finding the tenth dwelling. Each field enumerator was assigned to only one polling booth, and hence the paths taken during each selection crossed each household only once, increasing the likelihood of a random and unbiased sample. Once a household is selected, a randomly chosen adult member (ages 18-60) of the household was chosen to answer our survey questions after they qualified based on pre-conditions.

The three pre-conditions of the survey were (1) access to a personal smartphone (i.e. not a shared household cellphone), (2) connectivity of the phone to working Internet for the past 6 months, (3) usage of WhatsApp on the phone.

These conditions ensured that access to WhatsApp and other social media accounts was by the respondent alone, and these restrictions were put into place to ensure that respondents in the study were likely to be exposed to political fake news over WhatsApp in the months leading up to the election. Sharing mobile phones is especially common among adults in semi-urban and rural India. Further, it is also more common for women than it is for men. Pew survey data from 2019 finds that women are less likely than men to own their own mobile phones, and consequently, significantly more women (20%) than men (5%) report sharing a device with someone else.

These sampling conditions resulted in an uneven age distribution for the study, with about 35% of respondents below age 22 and only about 6% of the sample above age 45. It also resulted in an uneven gender distribution. Focus group discussions with men and women above the age of 45 showed that people in this age group largely did not own their own cellphones; they

reported having shared cellphones used by the entire house or not having access to a phone with working Internet at all. Women, particularly, reported using their husbands' cellphones to communicate and did not report owning their own social media accounts. As a result, only 6 of the women in this sample were above the age of 40.

C Flyers

Respondents were given flyers as part of the intervention. For treatment group respondents, the front side of the flyer included four false political stories that went viral on social media in the months before the 2019 election. The flyer included the photos / screen grabs associated with these fake stories along with an explanation for what the correct version of the story is. The back of the flyer contained 6 general tips to spot misinformation. Enumerators explained each bit of information in the flyer and then finally handed the flyers over to respondents. Treatment 1 flyer has pro-BJP fake stories, Treatment 2 flyer has anti-BJP stories, the control flyer is a placebo and has information on plastic pollution. All materials were in Hindi (English translations below).

Figure C.1: Treatment 1 – Pro-BJP Flyer (front and back)

Figure C.2: Treatment 2 – Anti-BJP Flyer (front and back)

Figure C.3: Placebo Control Flyer (front and back)

Together we can fight plastic usage!

Plastic is harmful for the environment. Here are some ways in which plastic causes damage.

Plastic never goes away. Plastic waste — whether in a river, an ocean, or on land — can persist in the environment for centuries. Most plastic items never fully disappear; they just get smaller and smaller.

Plastic bags cause a serious danger to birds and marine animals that mistake them for food. More than a million birds die each year from plastic pollution. Plastic particles are also swallowed by farm animals or fish.

Plastic affects human health. Because animals mistake it for food, plastic can find their way onto our dinner plates. Exposure to plastic is linked to a number of diseases and ailments.

By clogging sewers and providing breeding grounds for mosquitoes and pests, plastic waste — especially plastic bags — can increase the transmission of vector-borne diseases like malaria.

TIPS TO REDUCE PLASTIC USAGE

Here are easy tips that you can implement to reduce plastic pollution.

1. **Don't use disposable packaging.** Avoid using disposable or single-use plastic. Examples of this are Bisleri water bottles, plastic straws, polythene bags.
2. **Always bring a bag to the shop.** Examples of alternatives to plastic are cloth, jute, and paper.
3. **Reuse.** If you do happen to be in possession of a plastic bag, make sure it is reused for different purposes. Before throwing plastic items, it is important to consider how they can be reused.
4. **Replace plasticware.** Plastic boxes, forks, spoons, plates can be replaced with steel and other materials.
5. **Do not litter.** If you want to throw away plastic, ensure that it is thrown in a bin or trash can. Do not discard plastic into water sources, rivers, or simply out in the open.
6. **Support bans.** Many municipalities in India have enacted bans on single use plastic bags, takeout containers, and bottles. You can support the adoption of such policies in your community.
7. **Spread the word.** Stop others from littering. Stay informed on issues related to plastic pollution and help make others aware of the problem.

This project is sponsored by the Center for the Advanced Study of India at the University of Pennsylvania

D Dependent Variables

To measure key outcomes of interest respondents were shown a series of fourteen news stories. These stories varied in content, salience, and critically, partisan slant. Half of the stories were pro-BJP in nature and the other half anti-BJP. Each respondent saw all the fourteen stories, but the order in which they were shown was randomized. Table D.1 lists the fourteen stories shown to respondents. Following each story, two primary dependent variables were measured:

1. Perceived accuracy of fake news identification, with the question "Do you believe this news story is fake?" (binary response, 1 if yes, 0 otherwise)

2. Confidence in identification of the story as fake or real, with the question "How confident are you that the story is real / fake?" (4-point scale, 1 = very confident, 4 = not confident at all)

Table D.1: Dependent Variable Stories

	Story	**Party Slant**	**Veracity**
1	Cow urine cures cancer	Pro-BJP	False
2	Photos of militant bloodshed in Kashmir w/ pro-army message	Pro-BJP	False
3	India has not experienced a single foreign terror attack since 2014	Pro-BJP	False
4	Photoshopped image of war hero in BJP attire	Pro-BJP	False
5	Images of the Indian flag projected onto the Statue of Liberty	Pro-BJP	False
6	Rumor that new Indian notes have tracking chips embedded	Pro-BJP	False
7	Rumor that the govt. has installed CCTV cameras in voting booths	Anti-BJP	False
8	Photoshopped images of BJP workers littering the Ganga river	Anti-BJP	False
9	Rumor that BJP workers use duplicate votes to rig elections	Anti-BJP	False
10	Rumors on lack of policing by govt. leading to child kidnapping	Anti-BJP	False
11	Photoshopped image of govt. built Patel statue developing cracks	Anti-BJP	False
12	Rumors of BJP voters hacking voting machines to rig elections	Anti-BJP	False
13	PM Modi has a new radio show on air called Mann Ki Baat	Pro-BJP	True
14	A recent attack killed 40 Indian CRPF soldiers in Kashmir's Pulwama	Anti-BJP	True

After the fourteen political stories, two additional dependent variables were measured: self-perceived efficacy of the treatment, and self-reported media literacy. Self-perceived efficacy was measured by asking respondents "How confident are you that you can spot fake news from real news?" (4-point scale, 1 = very confident, 4 = not confident at all). Media literacy was measured in two ways: trust in news received over WhatsApp (4-point scale); and how frequently they forwarded political messages over WhatsApp (6-point scale). Self-reported literacy and efficacy

were measured to determine whether the intervention was successful at generating awareness of the problem of misinformation, arguably demonstrated by decreased trust in WhatsApp and forwarding of political stories. Finally, voter turnout was measured. This was done by asking respondents to show enumerators the index finger of their left hand, which, if they voted, would be marked with purple indelible ink. Because respondents were surveyed within a few days of having voted, the presence of an inked finger is a clean and perfect measure of voter turnout. Though this may not be true for instances where respondents refuse to show their ink, in this study every respondent willingly showed enumerators their index finger and no one refused.

Table D.2: ATE and HTE for Confidence in Story Veracity

	Dependent variable: Confidence in Story Veracity		
	Average Confidence Level		
	(1)	(2)	(3)
Literacy Intervention	−0.006	0.058***	−0.045**
	(0.006)	(0.022)	(0.020)
Education		0.003*	
		(0.002)	
Male			−0.020
			(0.017)
Literacy Intervention × Education		−0.007***	
		(0.002)	
Literacy Intervention × Male			0.044**
			(0.021)
Constant	0.937***	0.875***	0.924***
	(0.005)	(0.018)	(0.016)
Observations	1,224	1,224	1,224
R^2	0.001	0.070	0.066
Adjusted R^2	−0.00004	0.066	0.062
Residual Std. Error	0.103 (df = 1222)	0.100 (df = 1218)	0.100 (df = 1218)
F Statistic	0.954	18.340***	17.181***
Note:			*p<0.1; **p<0.05; ***p<0.01

The analysis in Table D.2 measures the effect of the treatment on self-reported confidence that respondents had in each story being true or false. Confidence was measured on a four-point scale between 0 and 1 for each story with higher numbers indicating more expressed confidence. The dependent variable was calculated as the average confidence level across all stories. While there is no main effect of the treatment on confidence, there is an effect with certain subgroups. Respondents who were more educated and received the intervention were significantly less likely to be confident in their responses. By contrast, men who received the intervention were more

likely to be confident in their responses relative to women.

Tables below identify the effect of the intervention on secondary dependent variables measured for this study. The first column estimates the effect of the intervention on self-reported confidence in being able to tell the difference between true and fake stories, that is, this measures the efficacy of the treatment. Confidence was measured on a three point scale where higher values indicate a greater level of confidence. In Column 2, the dependent variable is self-reported scrutiny of messages; respondents were asked whether they check if messages are true before forwarding them. This is a binary variable. In Column 3, respondents' turnout in the general election is measured. In the final column, I measure trust in WhatsApp on a four-point scale where higher values indicate more trust in the medium.

Table D.3 is the average treatment effect on the four dependent variables described above. Table D.4 is the heterogeneous effect of party identity on the four dependent variables described above.

Table D.3: Average Treatment Effect on Non-Identification DVs

	Dependent variable:			
	Confidence	Message Checking	Turnout	WhatsApp Trust
	(1)	(2)	(3)	(4)
Literacy Intervention	0.001	−0.015	−0.013	−0.041
	(0.023)	(0.026)	(0.030)	(0.040)
Constant	0.170***	0.246***	0.478***	2.539***
	(0.019)	(0.021)	(0.025)	(0.033)
Observations	1,224	1,224	1,224	1,224
R^2	0.00000	0.0003	0.0002	0.001
Adjusted R^2	−0.001	−0.001	−0.001	0.00004
Residual Std. Error (df = 1222)	0.377	0.425	0.499	0.663
F Statistic (df = 1; 1222)	0.003	0.350	0.192	1.051
Note:				*p<0.1; **p<0.05; ***p<0.01

Table D.4: Heterogeneous Effect of Party on Non-Identification DVs

	Dependent variable:			
	Confidence	Message Checking	Turnout	WhatsApp Trust
	(1)	(2)	(3)	(4)
Literacy Intervention	−0.025	−0.016	−0.038	0.009
	(0.041)	(0.046)	(0.054)	(0.071)
BJP Supporter	0.012	−0.022	0.035	0.103
	(0.040)	(0.045)	(0.053)	(0.070)
Literacy Intervention x	0.039	0.002	0.035	−0.075
BJP Supporter	(0.049)	(0.055)	(0.065)	(0.086)
Constant	0.162***	0.262***	0.454***	2.469***
	(0.033)	(0.037)	(0.044)	(0.058)
Observations	1,224	1,224	1,224	1,224
R^2	0.003	0.001	0.003	0.003
Adjusted R^2	0.0003	−0.002	0.001	0.0004
Residual Std. Error (df = 1220)	0.376	0.425	0.499	0.663
F Statistic (df = 3; 1220)	1.111	0.335	1.377	1.175
Note:				*p<0.1; **p<0.05; ***p<0.01

E Enumerator Fixed Effects

The endline survey to measure the dependent variable was conducted offline (as a paper survey) for field safety reasons. The main dependent variable consisted of 14 stories, but because the survey was conducted offline, the order of appearance of these stories was pre-determined and limited to 3 random orders. A single enumerator only had access to one of the three random orders. Hence as a robustness check, I replicate the main results with enumerator fixed effects.

Table E.1 replicates results for the main effect of the intervention on the outcome. Results are robust to enumerator fixed effects.

Table E.1: Effect of Treatment with Enumerator Fixed Effects

	Dependent variable: Number of Stories Identified as Fake			
	Pro-BJP Stories	Anti-BJP Stories	Pro-BJP Stories	Anti-BJP Stories
	(1)	(2)	(3)	(4)
Literacy Intervention	−0.007	−0.004		
	(0.058)	(0.053)		
Literacy + Pro-BJP Fact-Check			0.003	0.001
			(0.067)	(0.061)
Literacy + Anti-BJP Fact-Check			−0.017	−0.008
			(0.067)	(0.061)
Constant	4.789***	5.741***	4.789***	5.741***
	(0.060)	(0.054)	(0.060)	(0.054)
Observations	1,224	1,224	1,224	1,224
R^2	0.252	0.123	0.252	0.123
Adjusted R^2	0.250	0.120	0.249	0.120
Residual Std. Error	0.961 (df = 1220)	0.868 (df = 1220)	0.962 (df = 1219)	0.868 (df = 1219)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table E.2 replicates results with enumerator fixed effects for the heterogeneous effect of party identity. Results are robust to enumerator fixed effects.

Table E.2: Effect of Treatment x Party with Enumerator Fixed Effects

	Dependent variable: Number of Stories Identified as Fake	
	Pro-BJP Stories	Anti-BJP Stories
	(1)	(2)
Literacy Intervention	0.254**	0.077
	(0.103)	(0.093)
BJP Supporter	0.265***	0.327***
	(0.102)	(0.092)
Literacy Intervention x	−0.384***	−0.120
BJP Supporter	(0.125)	(0.112)
Constant	4.608***	5.521***
	(0.092)	(0.082)
Observations	1,224	1,224
R^2	0.258	0.139
Adjusted R^2	0.255	0.135
Residual Std. Error (df = 1218)	0.958	0.860
F Statistic (df = 5; 1218)	84.543***	39.252***
Note:		*p<0.1; **p<0.05; ***p<0.01

F All Stories As DV

Below I replicate results where the dependent variable is the number of stories correctly identified given all fourteen stories, true and false. Results hold.

Table F.1: Effect of Treatment on Identification of Fake News

	Dependent variable: Number of Stories Accurately Identified	
	(1)	(2)
Literacy Intervention Pooled	−0.005	
	(0.097)	
Literacy + Pro-BJP Fact-Check		0.014
		(0.112)
Literacy + Anti-BJP Fact-Check		−0.024
		(0.113)
Constant	11.638***	11.638***
	(0.080)	(0.080)
Observations	1,224	1,224
R^2	0.00000	0.0001
Adjusted R^2	−0.001	−0.002
Residual Std. Error	1.604 (df = 1222)	1.605 (df = 1221)
F Statistic	0.002 (df = 1; 1222)	0.058 (df = 2; 1221)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table F.2: Effect of Treatment × Party on Identification of Fake News

	Dependent variable: Number of Stories Identified as Fake
	(1)
Literacy Intervention	0.400**
	(0.172)
BJP Supporter	0.497***
	(0.170)
Literacy Intervention ×	−0.595***
BJP Supporter	(0.208)
Constant	11.300***
	(0.140)
Observations	1,224
R^2	0.007
Adjusted R^2	0.005
Residual Std. Error	1.599 (df = 1220)
F Statistic	3.067** (df = 3; 1220)
Note:	*p<0.1; **p<0.05; ***p<0.01

G Correlates of Misinformation

Independent of the literacy intervention, it is descriptively interesting for the understudied context of India to understand who is more likely to consume fake news and more likely to be able to identify news as false. I consider the main effect of several demographic and pre-treatment variables on ability to identify fake news. The results are presented in Table G.1. For all 12 dependent variable stories taken together, BJP partisans are significantly better at identifying fake news as compared to their non-BJP partisan counterparts. Further, as expected, accurate prior beliefs about fake stories are more likely to aid in identifying fake news. Higher levels of digital literacy were negatively associated with fake news identification, underscoring that greater knowledge of WhatsApp leads to more vulnerability to fake news in this context. However, those who report using WhatsApp more often are more likely to be able to identify fake news. Interestingly, higher levels of trust in WhatsApp do not correlate with identification of fake news stories, suggesting that familiarity with the medium itself can make people more vulnerable to misinformation and consequently more likely to share fake news.

With respect to demographic variables, increase in age is associated with a higher capacity to identify fake news. On the other hand, education has a positive effect on ability to identify news as fake.

Table G.1: Main Effect of Demographic and Pre-Treatment Variables

	Dependent variable: Number of Stories Identified as Fake
	Pooled DV : All Stories
Literacy Intervention	-0.060
	(0.095)
BJP Supporter	0.234^{**}
	(0.113)
Accurate Priors	0.480^{**}
(Higher = more accurate)	(0.231)
Digital Literacy	-1.168^{***}
(Higher = more literate)	(0.252)
Political Knowedge	-0.070
(Higher = more knowledge)	(0.046)
WhatsApp Use Frequency	0.150^{***}
(Higher = more usage)	(0.047)
Trust in WhatsApp	-0.071
(Higher = more trust)	(0.057)
Education	0.045^{**}
	(0.018)
Age	0.022^{***}
	(0.005)
Male	0.164
	(0.164)
Hindu	-0.185
	(0.144)
Constant	8.987^{***}
	(0.437)
Observations	1,158
R^2	0.066
Adjusted R^2	0.057
Residual Std. Error	1.509 (df = 1146)
F Statistic	7.335^{***} (df = 11; 1146)
Note:	*p<0.1; **p<0.05; ***p<0.01

H Age and Digital Literacy

I explore further the relationship between age, fake news identification, and digital literacy. The tables below look at age as variable. In Table H.1, I demonstrate that older respondents are better at identification. However in Table H.2, I find that older respondents have lower levels of digital literacy, demonstrating that despite having better digital literacy skills, younger respondents are worse are identifying fake news.

Table H.1: Effect of Age on Fake News Identification

	Dependent variable: Number of Stories Identified As Fake
	(1)
Age (Continuous)	0.024***
	(0.005)
Constant	9.276***
	(0.136)
Observations	1,224
R^2	0.019
Adjusted R^2	0.019
Residual Std. Error	1.553 (df = 1222)
F Statistic	24.246*** (df = 1; 1222)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table H.2: Effect of Age on Digital Literacy

	Dependent variable: Digital Literacy (Higher = More Literate)
	(1)
Age (Continuous)	−0.001**
	(0.001)
Constant	0.796***
	(0.017)
Observations	1,224
R^2	0.005
Adjusted R^2	0.004
Residual Std. Error	0.194 (df = 1222)
F Statistic	5.716** (df = 1; 1222)
Note:	*p<0.1; **p<0.05; ***p<0.01