Educative Interventions to Combat Misinformation:

Evidence From a Field Experiment in India*

Sumitra Badrinathan†
University of Pennsylvania

Word Count: 11,846

Abstract

Misinformation makes democratic governance harder, especially in developing countries. Despite its real-world import, little is known about how to combat misinformation 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 misinformation during the 2019 elections (N=1224). Treated respondents received an hour-long in-person media literacy training in which enumerators demonstrated two tools to identify misinformation, provided corrections to stories, and discussed the importance of verifying information. Receiving this hour-long media literacy intervention did not significantly increase respondents' ability to identify misinformation on average. However, treated respondents who support the ruling party became significantly less able to identify pro-attitudinal stories. 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 Guy Grossman, Matthew Levendusky, Marc Meredith, Devesh Kapur, Neelanjan Sircar, Brendan Nyhan, Tariq Thachil, Douglas Ahler, Milan Vaishnav, Adam Ziegfeld, Devra Moehler, Jeremy Springman, Emmerich Davies, Simon Chauchard, Simone Dietrich and anonymous referees. 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, the Harvard Experimental Political Science Conference, APSA 2020, SPSA 2020, and the UCLA Politics of Order and Development lab.

†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. 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 rumors are as extreme as violence, demonstrating that misinformation 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? Nearly all of the extant literature on combating misinformation focuses on the U.S. and other developed democracies, where misinformation spreads via public sites such as Facebook and Twitter. Interventions in these contexts are not easily adapted for misinformation distributed on encrypted chat applications such as WhatsApp, where no one, including the app developers themselves, can see, read or analyze messages. Encryption necessitates that the the burden of fact-checking fall solely on the user and therefore, the more appropriate solutions in such contexts are bottom-up, user-driven learning and fact-checking to combat misinformation.

This study is one such bottom-up effort to counter misinformation 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 misinformation in India. To answer this question, I implemented a large-scale field experiment with 1,224 respondents in the state of Bihar in India during the 2019 general elections, when misinformation was arguably at its peak. In an hour-long intervention, treatment group respondents were taught concrete tools to verify information (for example, performing reverse image searches, navigating a fact-checking website) along with a corrective intervention to false stories. After a two-week period, respondent households were revisited to measure their ability to identify misinformation.

My experiment shows that an hour-long, educative treatment is not sufficient to help respondents combat misinformation. Importantly, the average treatment effect is not significantly distinguishable from zero. Finding that an in-person, hour-long and bottom-up learning intervention does not move people's prior attitudes is testimony to the tenacity and destructive effects of misinformation in low education settings such as India. 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 subgroups. Bharatiya Janata Party (BJP) partisans (those self-identifying as supporters of the BJP, the national right-leaning party in India) who receive the treatment are *less* likely to identify pro-attitudinal stories as false. 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, at least with BJP partisans and during elections.

This study hopes to spark a research agenda on the ways to create an informed citizenry in low income democracies through testing and implementation of bottom-up measures to fight misinformation. 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 Misinformation and How Do We Fight it?

I define misinformation as claims that contradict or distort common understanding of verifiable facts (Guess and Lyons 2020) and fabrications that are low in facticity (Tandoc Jr, Lim, and Ling 2018; Li 2020).

The literature on misinformation identifies three key components of false stories: 1) low levels of facticity, 2) journalistic presentation, and 3) intent to deceive (Egelhofer and Lecheler 2019; Farkas and Schou 2018). Given my focus on misinformation in India, my definition does not include the format of the news. In India, misinformation is spread via WhatsApp where much of it is in the form of text messages, with the content copied and pasted into the body of the message, such that it exists as standalone content. Hence this cannot mimic legitimate news websites and is rarely presented in a journalistic format. Further, while the creation of falsehoods in the Indian context can stem from organized attempts by political parties with the intention to deceive, users in WhatsApp groups who are the victims of such campaigns may further propagate falsehoods inadvertently or unintentionally. Thus my definition also leaves out the intention to deceive, defined in the literature as "disinformation" (Tucker et al. 2018).

A predominant model of misinformation 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. 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). Thus individuals' preexisting beliefs strongly affect their responses to corrections (Flynn, Nyhan, and Reifler 2017). Importantly, a number of contextual and individual moderators of such motivated reasoning predispose subsets of the population to be more vulnerable to misinformation.

The two key political factors that contribute to the vulnerability to misinformation effects are political sophistication and ideology (Wittenberg and Berinsky 2020). More politically sophisticated individuals (including political knowledge and education) are more likely to be resistant to corrections (Valenzuela et al. 2019) and are the least amenable to updating beliefs when misinformation supports their existing worldviews (Lodge and Taber 2013). Additionally, differences in informedness can affect how well corrective information helps individuals develop knowledge about current events: Li and Wagner (2020) find that uninformed individuals are more likely to update their beliefs than misinformed individuals after exposure to corrective information. Further, ideology and partisanship are associated with differences in response to corrections. Although everyone is vulnerable to misinformation to a certain extent, worldview backfire effects are more visible for Republicans but not Democrats, given that the insular nature of the conservative media system is more conducive to the spread of misinformation (Faris et al. 2017; Ecker and Ang 2019; Nyhan and Reifler 2010).

Apart from political factors, research highlights age as a key demographic variable influencing both exposure to misinformation as well as responses to it. Studies find that older adults are more likely than others to share misinformation (Grinberg et al. 2019) and that the relationship between age and vulnerability to misinformation persists even after controlling for partisanship and ideology.

But despite the growing attention to misinformation in media and scholarship, empirical literature finds that the online audience for misinformation is a small subset of the total online audience. Those consuming false stories are a small, disloyal group of heavy internet users (Nelson and Taneja 2018): Grinberg et al. (2019) find that one percent of twitter users in their sample account for 80% of misinformation exposures; Guess, Nyhan, and Reifler (2018) find that almost 6 in 10 visits to fake websites came from the 10% of people with the most conservative online information diets. However, though people online are not clamoring for a continuous stream of false stories, misinformation in a multi-faceted and fast paced online environment can command people's limited attention (Guess and Lyons 2020). 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. These interventions can be grouped into reactive or top-down interventions that are implemented after misinformation is seen, and proactive or bottom-up interventions that seek to fight misinformation before it has been encountered.

Examples of top-down interventions 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). 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.

Bottom-up interventions to combat misinformation rely on inoculation theory, the idea of preparing people for potential misinformation by exposing logical fallacies inherent in misleading communications a priori (Compton 2013). To this end, Tully, Vraga, and Bode (2020) and Vraga, Bode, and Tully (2020) conduct experiments where treatment group respondents were reminded to be critical consumers of the news via tweets encouraging people to distinguish between high- and low-quality news. Wineburg and McGrew (2019) demonstrate that lateral reading (i.e. cross-checking information with additional sources) led to more warranted conclusions relative to vertical reading (i.e. staying within a website to evaluate its reliability). Cook, Lewandowsky, and Ecker (2017) inoculated respondents against misinformation by presenting mainstream scientific views alongside contrarian views.

Closer in design to the present study, Guess et al. (2020) evaluate a digital literacy intervention in India and the United states utilizing the "tips" provided by WhatsApp to measure whether they are effective at increasing the perceived accuracy of true stories. In Hameleers (2020), similar tips to spot misinformation are paired with fact checks in a bundled treatment.

Moving beyond the Western context, a small but burgeoning literature looks at misinformation on WhatsApp in developing countries. In Brazil, Rossini et al. (2020) compare misinformation sharing dynamics on WhatsApp and Facebook in Brazil and find that those who are more engaged in political talk are significantly more likely to have shared misinformation. In Zimbabwe, Bowles, Larreguy, and Liu (2020) use WhatsApp to disseminate messages aimed at targeting COVID-19 misinformation and find that intervention resulted in significantly greater knowledge about best practices. In Indonesia, Mujani and Kuipers (2020) find that younger, better-educated, and wealthier voters were more likely to believe misinformation. Finally, in India, using interview data with stakeholders and focus groups, Banaji et al. (2019) describe a misinformation ecosystem on WhatsApp that is built to exploit ideologies, creating rifts between social groups and paving the way for misinformation to result in violence and increased prejudice. Additionally, Garimella and Eckles (2020) use machine learning algorithms to analyze WhatsApp messages in India and determine that image misinformation is highly prevalent on public WhatsApp groups.

Apart from user-driven corrections to misinformation, two studies suggest platform-based or algorithmic solutions. Using data from public WhatsApp groups in India and Brazil, Reis et al. (2020) find that a significant proportion of misinformation on these groups include stories that have already been fact-checked on other platforms and thus propose a solution where WhatsApp can detect when a user shares content that has been previously labeled by fact-checkers as misinformation. Relatedly, Melo et al. (2019) create a public tool to monitor content on WhatsApp groups, including most-shared stories and categories of content, and apply the tool to monitor public misinformation during the 2018 Brazilian general election.

Despite the inroads made in the study of misinformation in developing countries, the problem of fake news in these contexts continues to grow. 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).

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 misinformation circulating in India 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. Then BJP President Amit Shah underscored this during a public address in 2018: "We can keep making messages go viral, whether they are real or fake, sweet or sour" (Wire 2018). 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 misinformation in India. Four characteristics make their users 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 that vulnerability to misinformation is impacted by lower literacy and education.

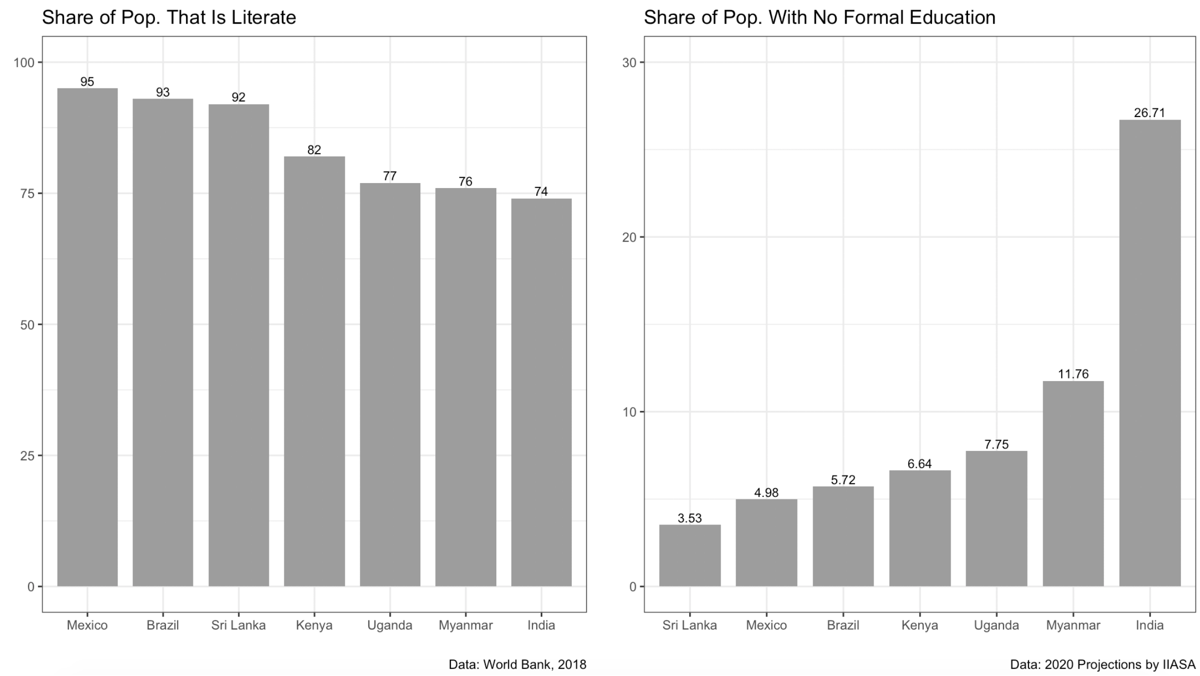


Figure 1: India Has Low Levels of Literacy and Education

Second, Internet access has exploded in the developing world. India, particularly, is digitizing faster than most mature and emerging economies, 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 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).

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). Paradoxically, this leap in development coupled with the novelty and unfamiliarity with the Internet could make new users more vulnerable to information received online. The example of Geeta highlights this aspect. Geeta lives in Arrah, Bihar 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 misinformation 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 misinformation. Critically, this means that top-down and platform-driven solutions are impractical in the case of private group chats on WhatsApp, suggesting that bottom-up interventions are more promising.

Finally, the format of misinformation in India is mainly visual: much of what goes viral on WhatsApp constitutes photo-shopped images and manufactured videos (Garimella and Eckles 2020). 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). On WhatsApp, false stories are almost never shared with a link; images are forwarded as is to thousands of users, making the original source unknown and difficult to trace.

Research on WhatsApp in India finds that a large fraction of groups belong to political parties or causes supporting political parties (Garimella and Eckles 2020). 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 politics 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.

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

Despite these findings, I argue that party identities will likely moderate attitudes in India. This is largely because of the nature of the BJP's appeals. The recent BJP administration under the leadership of Prime Minister Narendra Modi represents a departure from traditional models of voting behavior in India, highlighting that Modi's rule is a form of personal politics in which voters prefer to centralize political power in a strong leader, and trust the leader to make good decisions for the polity (Sircar 2020). Some have concluded that under Modi, polarization in India is more toxic than it has been in decades, showing no signs of abating (Sahoo 2020). To add to this, misinformation is India is inherently political in nature, with disinformation campaigns often stemming from party sources themselves (Singh 2019). Finally, partisan identities tend to be more salient during elections, when citizen attachments to parties are heightened (Michelitch and Utych 2018). Taken together, these three factors indicate that BJP partisans are more likely to respond expressively to the partisan treatment and engage in motivated reasoning in the face of counter-attitudinal information.

5 Media Literacy Intervention

I designed a pedagogical, in-person media literacy treatment with educational tools to address misinformation in the Indian context.

The concept of media literacy captures the skills and competencies that promote critical engagement with messages produced by the media, needed to successfully navigate a complex information ecosystem (Jones-Jang, Mortensen, and Liu 2019). Research finds that media literacy can bolster skepticism toward false and misleading information, making it particularly suitable to address the spread of misinformation (Kahne and Bowyer 2017). Experimental studies promoting media literacy initiatives against misinformation operationalize media literacy by increasing the salience of critical thinking (Vraga, Bode, and Tully 2020) or by gauging respondent knowledge about media industries and systems (Vraga and Tully 2019) or going a step further by providing tips to spot misinformation (Guess et al. 2020).

But simply nudging respondents to be more critical consumers or providing tips asking them to be more aware may be insufficient to help counter misinformation in contexts where respondents are not armed with the *tools* to apply such advice to the information they encounter. These tools can be thought of as a set of instructions needed to apply a tip, or as concrete steps which can be used to spot or correct misinformation.

In contexts such as India where respondents may be unconscious Internet users unaware about misinformation, tips (such as accuracy nudges) may not be sufficient to change behavior. Such nudges might prime the concept of accuracy, but if respondents do not know how to apply the information in those nudges, in spite of having the desire to, they may fall short in reaching their accuracy goals.

5.1 Experimental Design

The intervention targeted to treatment group respondents was, by design, a bundled treatment incorporating several elements, drawing on research demonstrating that the most promising tools to fight misinformation are fact-checking combined with media literacy (Hameleers 2020). The intervention consisted of surveying a respondent in their home and undertaking the following activities in a 45-60 minute visit:

1. Pre-treatment survey: Field enumerators administered survey modules to measure demographic and pre-treatment covariates including digital literacy, political knowledge, media trust, and prior beliefs about misinformation.[2]

2. Pedagogical intervention: Next, respondents learnt two concrete tools to identify misinformation.

   Performing reverse image searches: A large part of misinformation in India comprises of misleading 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. 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 discussed and demonstrated 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 involved demonstrating to respondents how to navigate a fact-checking website, www.altnews.in, a non-profit fact-checking service in India.

3. Corrections and tips flyer: Enumerators next helped respondents apply these tools to fact-check four false 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 false story, 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 misinformation. 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

[2]Summary statistics for all key variables are included in Table A.1

[3]Interview with Bharat, March 31, 2019.

flyer containing tips to reduce plastic usage.

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

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	Tools + corrections to 4 pro-BJP false stories
T2: Pedagogical Intervention + Anti-BJP flyer	Tools + corrections to 4 anti-BJP false stories
Control: Plastic Pollution Intervention + flyer	Tools + tips on plastic pollution

Respondents in both treatment groups received the pedagogical intervention. However, one group received corrections to four pro-BJP false stories, the other received corrections to four anti-BJP false 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 and were given a flyer on tips to reduce plastic usage. The false stories 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 false stories. 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. Those respondents who identified with the BJP were one block, those who identified with any other party 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 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.

Respondents were selected through a random walk procedure. Within the sampling area, a random sample of polling booths (smallest administrative units) were selected to serve as enumeration areas. Within each enumeration area, enumerators were instructed to survey 10-12 households following a random walk procedure. This method was chosen over traditional survey listing techniques so as to minimize enumerator time spent in the field during the elections as well as because of a lack of accurate census data for listing (Lupu and Michelitch 2018). 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 was selected, respondents could qualify for the study based on three pre-conditions designed to maximize familiarity with the Internet: respondents were required to have their own cellphone (i.e. not a shared household phone), working Internet for 6 months prior to the survey, and WhatsApp was required to be downloaded on the phone. If multiple members of a household qualified based on the pre-conditions, a randomly selected adult member was requested to participate in the study.

Of note, only 20% of all households sampled had respondents who met the criteria for recruitment into the study. In Bihar, where only 20-30% of citizens have access to the Internet, this is unsurprising. Despite this, the study had a high response rate: of all those who were eligible for the study, 94.5 percent agreed to participate. The final sample comprised of 1,224 respondents. [4]

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 2 summarizes the timeline for this study.

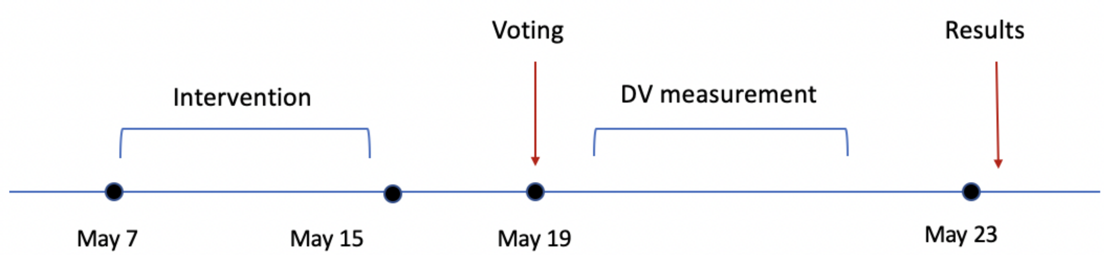


Figure 2: Experimental Timeline (May 2019)

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

[4]Additional details about the sampling process are available in Online Appendix B.

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.

5.3 Dependent Variables

In the endline survey, enumerators revisited respondents after they had voted. 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 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 misinformation. To this end, respondents were shown a series of fourteen news stories.[6] 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. [7] Each respondent saw all the fourteen stories, but the order in which they were shown was randomized.[8] A list of the fourteen stories shown to respondents is presented in Table D.1. Following each story, the key primary dependent variable measured perceived accuracy of stories, with the following question.[9]

"Do you believe this news story is false?" (binary response, 1 if yes, 0 otherwise)

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

[6] 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. More analysis of the true stories is in Online Appendix H.

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

[8] 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, E.2 and E.3.

[9] Online Appendix D describes secondary dependent variables measured.

5.4 Hypotheses and Estimation

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

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

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

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

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

In the equations, i represents the respondent, the *Intervention* variable in Equation 5.1 represents pooled assignment to the 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 *MisinformationId* counts the number of stories accurately identified as true or false. *MisinformationId* has been coded such that a positive estimated β_1 indicates an increase discernment ability.

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. In line with the literature on partisan motivated reasoning (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 false stories, or non-BJP partisans receiving corrections to pro-BJP false 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:

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

In Equation 5.3, *PartyID* is an indicator variable that takes on the value of 1 if the respondent 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 discernment ability among BJP partisans due to the treatment.

However, while partisanship might moderate attitudes, the role of other theoretical moderators such as age is unclear in the Indian context. Owing to lack of priors in the Indian context, I do not form pre-registered hypotheses about demographic moderators. Instead, I examine their relationships with misinformation through the following exploratory research questions:

RQ1: What is the relationship between age and vulnerability to misinformation? Does this relationship change as a function of the treatment?

RQ2: Does age correlate negatively with digital literacy, as in the American context? Are more digitally literate respondents likely to learn better from the treatment?

RQ3: What is the relationship between political sophistication (measured both by education and political knowledge) and vulnerability to misinformation?

6 Data and Results

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

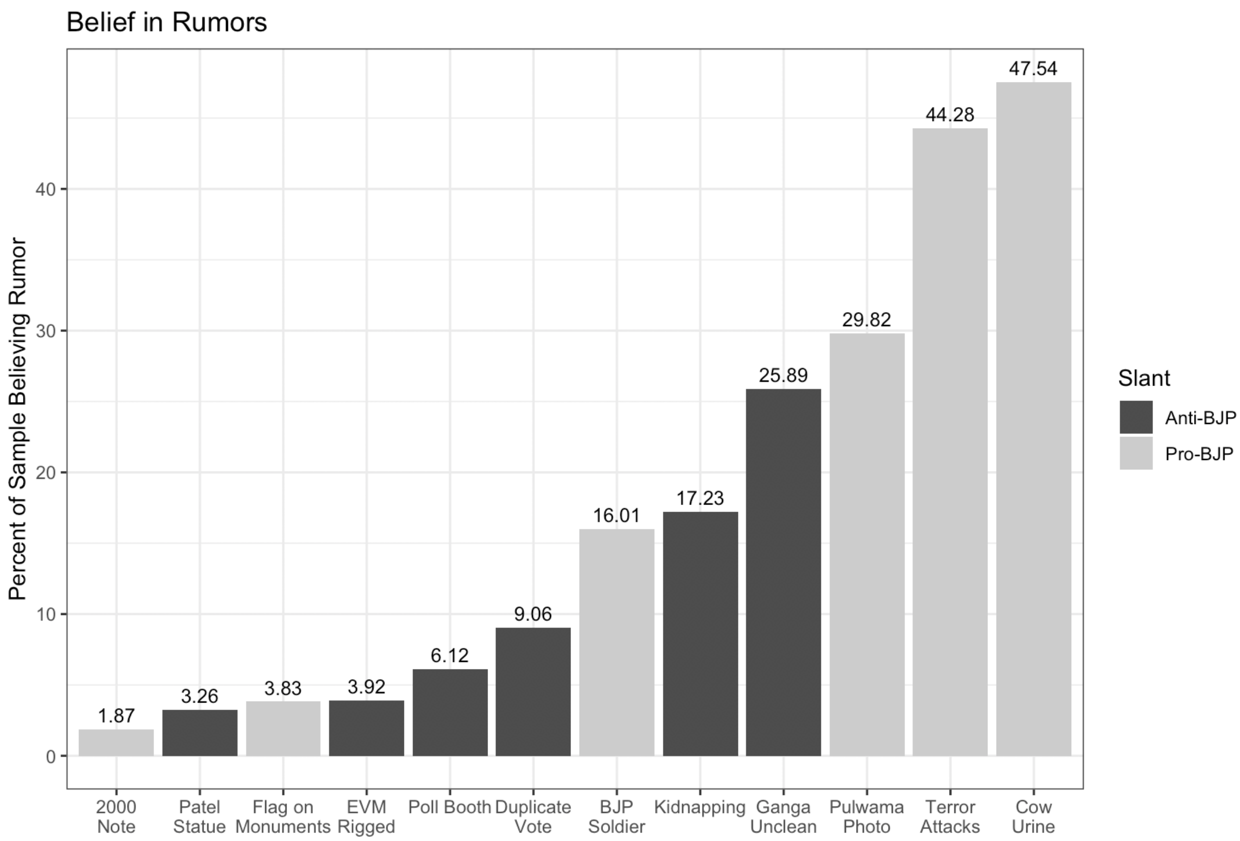
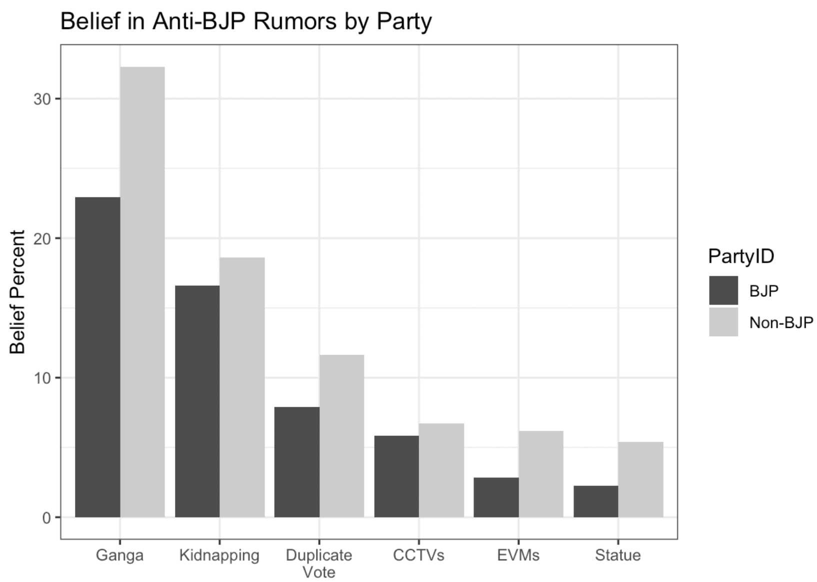


Figure 3: Percent of Sample Who Believe Rumors

Figure 3 lists the 12 false stories 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. Two 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 misinformation appears to be stronger, possibly alluding to its increased salience (Jerit and Barabas 2012), frequency of appearance on social media (Garimella and Eckles 2020), or to the presence of a higher proportion of BJP supporters in the sample. Overall, across the 12 rumors, respondents correctly classified an average of 9.91 rumors.

Figure 4 plots respondent belief in stories by partisan identity. For 10 out of the 12 partisan stories, we see a correspondence between respondent party identity and pre-tested political slant of the story. 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 an-



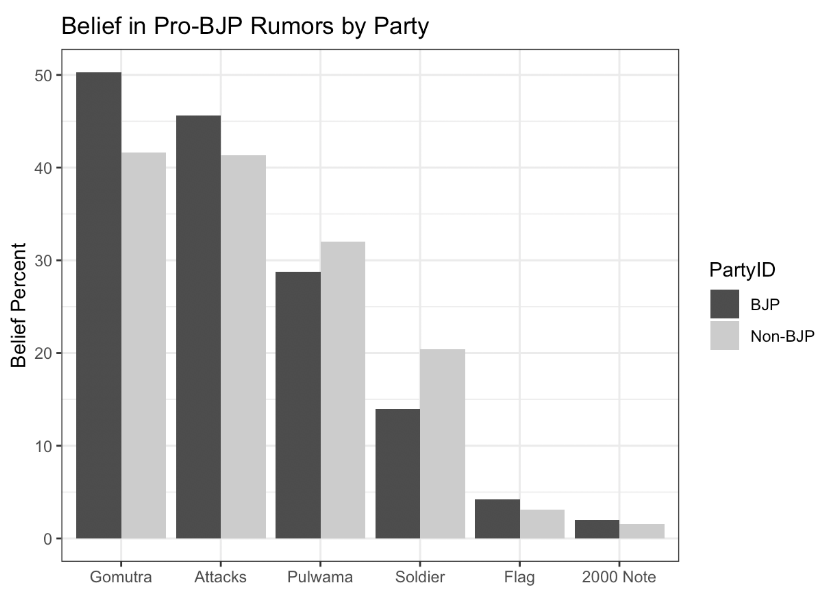


Figure 4: Belief in Rumors by Respondents' Party ID

alyze the correlates of misinformation among control group respondents (N=406). This analysis provides the baseline rate of identification ability in the absence of the intervention. In the regression analysis in Table 2, the dependent variable is the number of stories accurately classified as true or false by control group respondents.

First, we observe that BJP supporters were significantly better at discernment between true and false stories. 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 misinformation.

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 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.

Parsing this result further, it appears that BJP supporters' better discernment is driven by their ability to identify anti-BJP stories as fake. I find that while there is no difference in discernment of pro-BJP stories between BJP and non-BJP partisans, the true difference in means is significantly greater than 0 for anti-BJP stories, with BJP respondents correctly identifying these stories at a higher rate. Thus the finding that BJP respondents identify stories at a higher rate is driven only by their identification of anti-BJP messages as false.[10]

Next, we observe interesting findings with respect to age and digital literacy. While findings on misinformation in the United States suggest that older adults are most likely to engage with fake sources (Grinberg et al. 2019), this data demonstrates that older age is associated with better discernment between true and false stories, highlighting that the vulnerability of older adults to misinformation differs dependent on whether the outcome of interest is sharing behavior or perceived accuracy. The result that older adults are better at discernment may be attributed to findings that older adults may strategically decide to share information despite their beliefs, and that their vast knowledge base might help information processing despite cognitive declines (Brashier and Schacter 2020).

Further, I find that increases in digital literacy are associated with lower discernment, contrary to findings in the American context that people who are less digitally literate are more likely to fall for misinformation and clickbait (Munger et al. 2018). Finally, I find that while political knowledge is not correlated with misinformation identification, education is associated with significant increases in discernment.[11]

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

[10]See Online Appendix J for more analysis on this point.

[11]I explore these results further in Online Appendix F.

Table 2: Misinformation Identification in Control Group

	Dependent variable: Number of Stories Accurately Classified
BJP Supporter	0.562**
	(0.197)
Digital Literacy	−1.183**
(Higher = more literate)	(0.426)
Political Knowledge	−0.010
(Higher = more knowledge)	(0.080)
Frequency of WhatsApp Use	0.192*
(Higher = more usage)	(0.086)
Trust in WhatsApp	0.081
(Higher = more trust)	(0.103)
Education	0.071*
	(0.032)
Age	0.030**
	(0.009)
Male	0.372
	(0.282)
Hindu	−0.278
	(0.255)
Constant	9.270***
	(0.741)
Observations	406
R^2	0.091
Adjusted R^2	0.070
Residual Std. Error	1.558 (df = 396)
F Statistic	4.413*** (df = 9; 396)
Note:	*p<0.05; **p<0.01; ***p<0.001

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 false 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 false 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. 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 misinformation as a function of teaching respondents media literacy tools.

Results are shown in Table 3. The key dependent variable is the number of stories that a respondent correctly classified as true or false. 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 model (1). In model (2), I split the treatment into the pro-BJP corrections and the anti-BJP corrections (note both treatment conditions receive the same literacy intervention).

Table 3 demonstrates that the intervention did not increase misinformation 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 misinformation among respondents in Bihar, India. 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 misinformation in India. Priors about misinformation in this context appear resistant to change but, as I demonstrate below, this does not preclude moderating effects of partisan identity.

I now turn to the analysis of heterogeneous effects of partisan identity. Results are striking: I find that the interaction effect of the treatment on BJP partisans produces a negative effect on the ability to identify misinformation (Table 4).

This indicates that the treatment made BJP supporters worse at identifying misinformation. However, since the main dependent variables pools together all stories– true and false, pro-BJP and not– I now examine whether the negative coefficient on the interaction term obtains for pro-attitudinal and counter-attitudinal stories separately. In Table 5, I limit the dependent variable to the set of false stories. In Column 1 I estimate the effect of receiving the treatment for

Table 3: Effect of Treatment on Ability to Identify Misinformation

	Dependent variable: Number of Stories Accurately Classified	
	(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*** (0.080)	11.638*** (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 4: Effect of Treatment × Party (Discernment Measure)

	Dependent variable: Number of Stories Accurately Classified
	(1)
Literacy Intervention	0.400** (0.172)
BJP Supporter	0.497*** (0.170)
Literacy Intervention × BJP Supporter	−0.595*** (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

BJP supporters on ability to identify pro-BJP false stories, Column 2 does the same with anti-BJP false stories. The treatment variable for both models pools across receiving any treatment relative to control.

Table 5: Effect of Treatment x Party (False Stories)

	Dependent variable: Number of Stories Identified as False	
	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 show that 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 misinformation. For pro-BJP stories, the treatment effect for non-BJP supporters was 0.277, indicating that those who did not support the BJP and received the treatment identified an additional 0.277 stories. However, the treatment effect for BJP supporters was -0.135, indicating that those who supported the BJP and received the treatment identified 0.135 fewer stories.

Visualizing this interaction effect in Figure 5, where I graph the predicted values from the interaction model in Equation 5.3, it appears 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 false stories, while treatment group respondents who do not identify as BJP partisans show an increase in their ability to identify false stories. Thus the treatment was successful with non-BJP partisans, and backfired for BJP partisans. Importantly, these effects obtain only for the set of false stories that is pro-BJP in slant (implying that their corrections could be perceived as pro-attitudinal for non-BJP partisans). In Figure 6 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

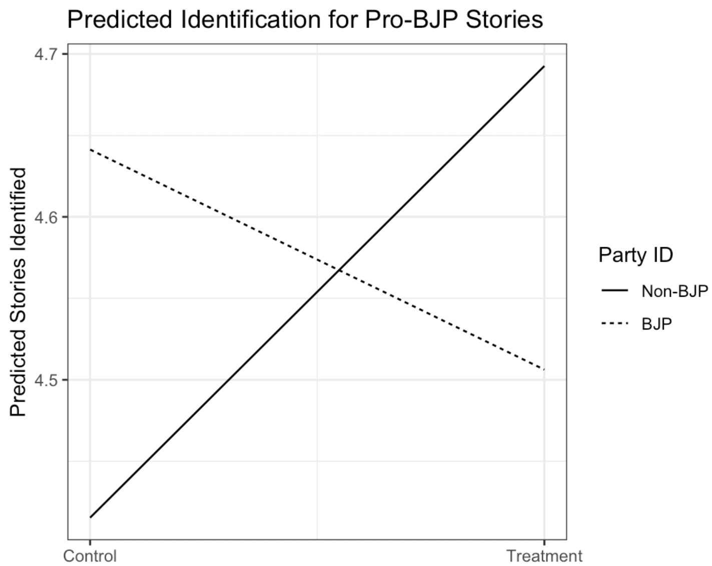


Figure 5: Predicted Identification of Pro-BJP Stories

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 false 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. 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. I examine this result further in the Discussion.

Moving beyond experimental results, I find that younger adults in the sample are less likely to be able to identify misinformation and that higher levels of digital literacy are associated with greater vulnerability to misinformation, contrary to findings in the United States (Munger et al. 2018; Grinberg et al. 2019). I also find that while political knowledge does not correlate with perceptions of stories, more educated respondents are better at spotting false stories. I explore these associations in Online Appendices F and G.

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). Paralleling results seen in developed contexts, the perceptual screen (Campbell et al. 1960) of BJP partisanship shaped how respondents interacted with this treatment, with BJP partisans demonstrating a tendency to cheerlead for their party and discredit pro-party stories despite them being false (Gerber and Huber 2009; Prior, Sood, and Khanna 2015). These

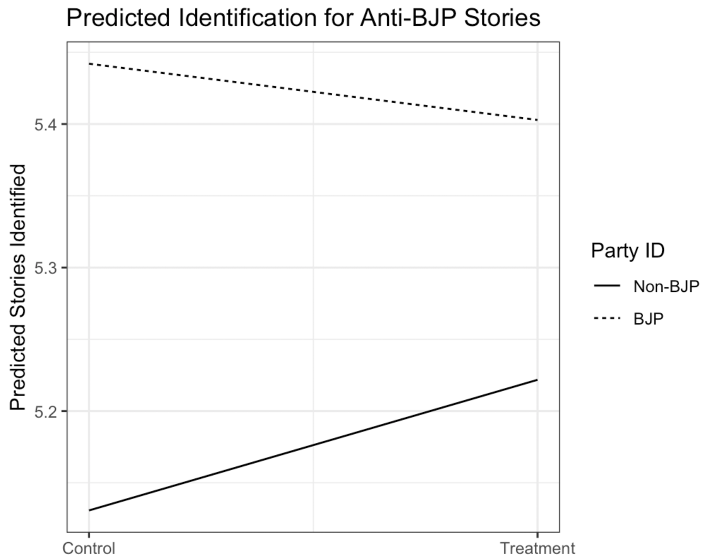


Figure 6: Predicted Identification of Anti-BJP Stories

findings of motivated reasoning demonstrate 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).

This finding is also surprising, given that there is little evidence of such backfire effects in the American context (Wood and Porter 2019). However, several other associations in the American context do not hold in this data: I find a positive correlation between increasing age and ability to discern true from false stories, a negative correlation between increasing digital literacy and discernment, and no association with political knowledge.[12] The nature of these findings underscores that what we know about misinformation comes largely from Western contexts and may not easily apply to other settings. It highlights that we need more theorizing and more data from non-Western contexts. Thus while I do find some backfire effects in this data, more needs to be done to establish the robustness of these findings. 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 between 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.

It is important to underscore that the intervention worsened misinformation identification only for the pro-BJP set of false stories. This effect does not appear for anti-BJP false stories. This highlights key differences in partisan identities in this data. First, though traditionally India has been described as a non-ideological system, the recent years under the Modi-led BJP

[12] See Online Appendix G for results.

governments have led some to conclude that tribalism and psychological attachments to political parties (Westwood et al. 2018) are more salient now than ever before (Sircar 2020). Importantly, such partisan attachments seem to have arisen in response to the personal popularity of Narendra Modi, with no comparable cult of personality on the political left. Thus it stands to reason that partisanship is stronger for BJP supporters. Second, the way the party identity variable is operationalized in the data further emphasizes this point. I cluster BJP supporters into one block and non-BJP supporters into another, but the non-BJP block is a heterogeneous group of respondents from several different parties. Thus we should expect that citizen attachments to political stories, true and false, will be perceived very differently for both political blocks. Third, political disinformation campaigns in India seem to emanate largely from the right-wing. This is underscored in my data by pro-BJP stories being believed to a much greater extent than anti-BJP stories, alluding to the fact that pro-BJP stories are more salient in the minds of respondents (Figure 3).

As a consequence of these factors, there is an inherent lack of symmetry between the two sets of stories that comprise the dependent variable measure. Pro-BJP stories are more salient and believed to a greater extent, hence there is likely more room for the treatment to move attitudes on the stories (as it does, for non-BJP supporters). On the contrary, the majority of anti-BJP stories were believed by less than 10% of the sample; this high ceiling might make it difficult for the treatment to work for anti-BJP stories.

7 Conclusion

Misinformation campaigns have the capacity to affect opinions and elections across the world. In this paper, I present new evidence on belief in popular misinformation stories in India in the context of the 2019 general elections. 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 misinformation 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 misinformation 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, but only for pro-attitudinal stories.

Despite these findings on partisanship, I consider some reasons why the average treatment effect was a null, along with some limitations of the study and future avenues for research. First, the two-week gap between the intervention and the measurement of outcomes is atypical for studies of this kind, where dependent variables are measured in close proximity to treatments. Thus is possible that a first-stage effect decayed over time and hence was not captured by the study. But while this design does not measure immediate effects, its results suggest that such treatments may have limited long-term durability.

Further, it is worth noting that this was an explicitly political intervention. Recent research demonstrates that news literacy messages paired with a political talk show host advancing an incongruent political argument increases the odds that people will rate the literacy message as biased (Tully and Vraga 2017). Thus, the political nature of the treatment itself likely activated motivated reasoning (Groenendyk and Krupnikov 2020).

Next, the timing of the intervention during a contentious election meant that not only were partisan identities more salient (Michelitch and Utych 2018) but also that the presence of several election officials, campaigning party workers, and GOTV efforts meant that respondents in the area had their door knocked on several times a day by different interest groups. Thus it is possible that the marginal effect of an additional house visit by the enumeration team for this study made the in-person intervention less salient.

Additionally, the design over-sampled false news stories in the outcome measure.[13] While this was done to maximise belief reduction in as many false stories with perilous consequences, systematically changing the balance of true and false stories to study how this factor shapes the efficacy of these types of campaigns.

Finally, while the outcome measured the perceived accuracy of news stories, it did not measure whether the participants used fact-checking tools between the intervention and the follow up. Future studies of this nature would benefit from test that allowed participants to directly apply the tools learned in the intervention at the moment of identification of false news stories. This would allow us to measure a number of intermediary steps including whether participants recalled the training and chose to apply it to identify misinformation. Thus a valuable prospect for future work would be to validate the usage and frequency of procedural tools before measuring beliefs.

But providing educative tools alone may be insufficient to help move attitudes, as media literacy is distinct from its application (Vraga, Bode, and Tully 2020). Several additions to such a design could be undertaken to ensure respondents are not only given tools, but also are motivated to apply them. As Tully, Vraga, and Bode (2020) note, the one-shot nature of such interventions may be insufficient in changing behavior in a fast-paced and overwhelming social media landscape. On social media, news literacy messages have to compete with many other pieces of information, and any one piece of information may get lost in the crowd. This may also be a factor in my study, as even though the intervention is fairly detailed, it is just a one-shot training session, and multiple in-person visits might be needed to reinforce these skills and provide respondents with the opportunity to practice them. One such strategy could be to make misinformation and digital literacy training a part of school curriculums. Further, highlighting

[13] While I specifically chose some false and some true stories—and explicitly told this to respondents—this might have made respondents more skeptical than they otherwise would have been, because we implicitly perceive most incoming information to be true (Brashier and Marsh 2020). Future studies can vary the ratio of true vs. false stories to examine the effects on the efficacy of the treatment. I thank an anonymous referee for highlighting this issue to me.

social responsibility might motivate citizens to use the tools at their disposal. Vraga, Bode, and Tully (2020) suggest that a more citizen-oriented framing might persuade people relative to a more personal framing, since it connects information to democratic duties. Similarly, Mullinix (2018) shows that heightening a sense of civic duty (i.e., citizens have an obligation to get the facts right) reduces partisan motivated reasoning, and a similar logic could apply here. Thus a related corollary would be to investigate whether accuracy motivations could also improve the applicability of these tools.

The findings from this study are local average treatment effects, dependent heavily on the locality where this experiment was conducted: the low-education, low-internet environment of semi-urban Bihar, at a time where politics was salient and where political misinformation was rife. New internet users, such as in this sample, not only face greater barriers in learning new technology, but the uptake of interventions designed to combat misinformation might be harder to implement with samples that are more rural and as a result have mobile-only and less stable internet connections. My results, along with those from similar studies (Guess et al. 2020) suggest that different techniques will be needed to reach diverse populations in the developing world. For better-educated, and more digitally literate populations, simpler tools may prove to be effective. But for new internet users, richer interventions that provide not only tools, but also boosts to motivation, may be necessary.

Thus, whether these findings would hold–or change–outside of this sample remains an open empirical question. Consequently, I caution about interpreting these null results to mean that interventions of this kind do not work, as thorough future work must look into replicating such a design in different contexts and times. Thus while this study was necessarily context-dependent, it is nevertheless an important first step towards tempering the human cost of misinformation in India.

References

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

Banaji, Shakuntala, Ramnath Bhat, Anushi Agarwal, Nihal Passanha, and Mukti Sadhana Pravin. 2019. "WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India." *Working Paper*. http://eprints.lse.ac.uk/104316/1/Banaji$_{w}hatsapp_{v}igilantes_{e}xploration_{o}f_{c}itizen_{r}eception_{p}ublished.pdf$.

Bowles, Jeremy, Horacio Larreguy, and Shelley Liu. 2020. "Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in Zimbabwe." *PloS one* 15 (10): e0240005.

Brashier, Nadia M, and Daniel L Schacter. 2020. "Aging in an era of fake news." *Current Directions in Psychological Science* 29 (3): 316–323.

Brashier, Nadia M, and Elizabeth J Marsh. 2020. "Judging truth." *Annual review of psychology* 71: 499–515.

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.

Compton, Josh. 2013. "Inoculation Theory." *The Sage handbook of persuasion: Developments in theory and practice* 2: 220–237.

Cook, John, Stephan Lewandowsky, and Ullrich KH Ecker. 2017. "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence." *PloS One* 12 (5): e0175799.

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/.

Ecker, Ullrich KH, and Li Chang Ang. 2019. "Political Attitudes and the Processing of Misinformation Corrections." *Political Psychology* 40 (2): 241–260.

Egelhofer, Jana Laura, and Sophie Lecheler. 2019. "Fake news as a two-dimensional phenomenon: a framework and research agenda." *Annals of the International Communication Association* 43 (2): 97–116.

Faris, Robert, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. "Partisanship, Propaganda, and Disinformation: Online media and the 2016 US Presidential Election." *Berkman Klein Center Research Publication* 6. https://bit.ly/3lmlcOl.

Farkas, Johan, and Jannick Schou. 2018. "Fake News as a Floating Signifier: Hegemony, Antagonism and the Politics of Falsehood." *Javnost-The Public* 25 (3): 298–314.

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.

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.

Garimella, Kiran, and Dean Eckles. 2020. "Images and misinformation in political groups: evidence from WhatsApp in India." *Working Paper*. https://arxiv.org/pdf/2005.09784.pdf.

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.

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.

Groenendyk, Eric, and Yanna Krupnikov. 2020. "What Motivates Reasoning? A Theory of Goal-Dependent Political Evaluation." *American Journal of Political Science*: 1–17.

Guess, Andrew, Brendan Nyhan, and Jason Reifler. 2018. "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign." *Working Paper*. http://www.ask-force.org/web/Fundamentalists/Guess-Selective-Exposure-to-Misinformation-Evidence-Presidential-Campaign-2018.pdf.

Guess, Andrew M, and Benjamin A Lyons. 2020. "Misinformation, Disinformation, and Online Propaganda." In *Social Media and Democracy: The State of the Field, Prospects for Reform*, ed. Nathaniel Persily and Joshua A Tucker. Cambridge University Press.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India." *Proceedings of the National Academy of Sciences* 117 (27): 15536–15545.

Hameleers, Michael. 2020. "Separating truth from lies: comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands." *Information, Communication & Society* 37 (2): 1–17.

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.

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.

Jones-Jang, S Mo, Tara Mortensen, and Jingjing Liu. 2019. "Does media literacy help identification of fake news? Information literacy helps, but other literacies don't." *American Behavioral Scientist* 00 (0): 1–18.

Kahne, Joseph, and Benjamin Bowyer. 2017. "Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation." *American Educational Research Journal* 54 (1): 3–34.

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.

Li, Jianing. 2020. "Toward a Research Agenda on Political Misinformation and Corrective Information." *Political Communication* 37 (1): 125–135.

Li, Jianing, and Michael W Wagner. 2020. "The Value of Not Knowing: Partisan Cue-Taking and Belief Updating of the Uninformed, the Ambiguous, and the Misinformed." *Journal of Communication* 70 (5): 646–669.

Lodge, Milton, and Charles S Taber. 2013. *The Rationalizing Voter*. Cambridge University Press.

Lupu, Noam, and Kristin Michelitch. 2018. "Advances in Survey Methods for the Developing World." *Annual Review of Political Science* 21: 195–214.

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.

Melo, Philipe, Johnnatan Messias, Gustavo Resende, Kiran Garimella, Jussara Almeida, and Fabrício Benevenuto. 2019. "Whatsapp monitor: A fact-checking system for whatsapp." In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13 pp. 676–677.

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/.

Mujani, Saiful, and Nicholas Kuipers. 2020. "Who Believed Misinformation during the 2019 Indonesian Election?" *Asian Survey* 60 (6): 1029–1043.

Mullinix, Kevin J. 2018. "Civic duty and political preference formation." *Political Research Quarterly* 71 (1): 199–214.

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.

Nelson, Jacob L, and Harsh Taneja. 2018. "The small, disloyal fake news audience: The role of audience availability in fake news consumption." *New Media & Society* 20 (10): 3720–3737.

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* (42): 939–960.

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

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.

Reis, Julio CS, Philipe Melo, Kiran Garimella, and Fabrício Benevenuto. 2020. "Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation?" *Harvard Kennedy School Misinformation Review*.

Rossini, Patrícia, Jennifer Stromer-Galley, Erica Anita Baptista, and Vanessa Veiga de Oliveira. 2020. "Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections." *New Media & Society*: 1461444820928059.

Sahoo, Niranjan. 2020. "Mounting Majoritarianism and Political Polarization in India." *Carnegie Endowment for International Peace*. https://carnegieendowment.org/2020/08/18/mounting-majoritarianism-and-political-polarization-in-india-pub-82434.

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.

Sircar, Neelanjan. 2020. "The politics of vishwas: political mobilization in the 2019 national election." *Contemporary South Asia* 28 (2): 178–194.

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.

Tandoc Jr, Edson C, Zheng Wei Lim, and Richard Ling. 2018. "Defining "Fake News": A typology of scholarly definitions." *Digital Journalism* 6 (2): 137–153.

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

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." *Hewlett Foundation report*. https://eprints.lse.ac.uk/87402/1/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf.

Tully, Melissa, and Emily K Vraga. 2017. "Effectiveness of a news media literacy advertisement in partisan versus nonpartisan online media contexts." *Journal of Broadcasting & Electronic Media* 61 (1): 144–162.

Tully, Melissa, Emily K Vraga, and Leticia Bode. 2020. "Designing and Testing News Literacy Messages for Social Media." *Mass Communication and Society* 23 (1): 22–46.

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.

Valenzuela, Sebastián, Daniel Halpern, James E Katz, and Juan Pablo Miranda. 2019. "The Paradox of Participation Versus Misinformation: Social Media, Political Engagement, and the Spread of Misinformation." *Digital Journalism* 7 (6): 802–823.

Vraga, Emily K, Leticia Bode, and Melissa Tully. 2020. "Creating News Literacy Messages to Enhance Expert Corrections of Misinformation on Twitter." *Communication Research* 00 (0): 1–23.

Vraga, Emily K, and Melissa Tully. 2019. "News literacy, social media behaviors, and skepticism toward information on social media." *Information, Communication & Society*: 1–17.

Westwood, Sean J, Shanto Iyengar, Stefaan Walgrave, Rafael Leonisio, Luis Miller, and Oliver Strijbis. 2018. "The tie that divides: Cross-national evidence of the primacy of partyism." *European Journal of Political Research* 57 (2): 333–354.

Wineburg, Sam, and Sarah McGrew. 2019. "Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information." *Teachers College Record* 121 (11): n11.

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.

Wittenberg, Chloe, and Adam J Berinsky. 2020. "Misinformation and Its Correction." In *Social Media and Democracy: The State of the Field, Prospects for Reform*, ed. Nathaniel Persily and Joshua A Tucker. Cambridge University Press.

Wood, Thomas, and Ethan Porter. 2019. "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence." *Political Behavior* 41 (1): 135–163.

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	68
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 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. It was also chosen over traditional listing methods due to lack of accurate census data and reliable addresses in the area.

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 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.

Importantly, these qualification conditions resulted in only 20% of all houses knocked on having a respondent who was eligible for the study. This is not atypical for Bihar, where only 20-30% of citizens have access to the internet. Despite this, the study also had a high response rate. Of all those who were eligible for the study, 94.5 percent agreed to participate. The high participation response rate corresponds to research in face-to-face surveys and in developing countries where response rates tend to be typically higher than in developed countries.

Of the 5.5% who refused, enumerator notes suggest that these respondents tended to be older women who (despite having a phone and internet) indicated they would be comfortable if the survey was conducted with a younger member of the household; in some cases they suggested enumerators wait inside the house until a younger member came back home. Once respondents consented to the survey and invited enumerators in their house, no respondent terminated the intervention early or asked that enumerators leave and come back at a different time. Thus, all respondents in the first wave who met the criteria and agreed to the survey completed the intervention in one setting.

The survey pre-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 misinformation 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 false 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 false stories, Treatment 2 flyer has anti-BJP false stories, the control flyer is a placebo and has information on plastic pollution. All materials were in Hindi and the survey and intervention were also administered in Hindi. Below I include English translations of the survey materials.

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)

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 news stories, with the question "Do you believe this news story is false?" (binary response, 1 if yes, 0 otherwise)

2. Confidence in identification of the story as false or real, with the question "How confident are you that the story is real / false?" (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 false 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 near-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.

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

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

self-reported confidence in being able to tell the difference between true and false 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 Accurately Classified	
	(1)	(2)
Literacy Intervention	−0.013	
	(0.084)	
Literacy + Pro-BJP Fact-Check		−0.005
		(0.097)
Literacy + Anti-BJP Fact-Check		−0.020
		(0.097)
Constant	12.144***	12.144***
	(0.087)	(0.087)
Observations	1,224	1,224
R^2	0.262	0.262
Adjusted R^2	0.260	0.260
Residual Std. Error	1.379 (df = 1220)	1.379 (df = 1219)
F Statistic	144.451*** (df = 3; 1220)	108.258*** (df = 4; 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 on a discernment measure of classifying true and false headlines. Results are robust to enumerator fixed effects. Table E.3 splits the key dependent variable into pro-BJP and anti-BJP stories.

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

	Dependent variable: Number of Stories Accurately Classified
Literacy Intervention	0.364^{**}
	(0.148)
BJP Supporter	0.555^{***}
	(0.146)
Literacy Intervention x BJP Supporter	-0.554^{***}
	(0.179)
Constant	11.768^{***}
	(0.131)
Observations	1,224
R^2	0.271
Adjusted R^2	0.268
Residual Std. Error	1.372 (df = 1218)
F Statistic	90.442^{***} (df = 5; 1218)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table E.3: Effect of Treatment x Party with Enumerator Fixed Effects

	Dependent variable: Number of Stories Identified as False	
	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 BJP Supporter	−0.384***	−0.120
	(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 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 misinformation 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 misinformation. The results are presented in Table F.1. For all 12 dependent variable stories taken together, BJP partisans are significantly better at identifying false stories as compared to their non-BJP partisan counterparts. Further, as expected, accurate prior beliefs are more likely to aid in identifying misinformation. However, higher levels of digital literacy were associated with worse levels of discernment, underscoring that greater knowledge of how to use WhatsApp may lead to more vulnerability to misinformation in this context. However, those who report using WhatsApp more often are more likely to be able to identify misinformation. Interestingly, higher levels of trust in WhatsApp do not correlate with identification of false stories, suggesting that familiarity with the medium itself can make people more vulnerable to misinformation and consequently more likely to share false stories.

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

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

	Dependent variable: Number of Stories Identified as False
	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

G Age and Digital Literacy

I explore further the relationship between age, misinformation, and digital literacy. The tables below look at age as variable. In Table G.1, I demonstrate that older respondents are better at discernment. However in Table G.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 false stories.

Table G.1: Effect of Age on Identification of Misinformation

	Dependent variable: Number of stories accurately classified
	(1)
Age (Continuous)	0.019^{***}
	(0.005)
BJP Supporter	0.242^{**}
	(0.115)
Digital Literacy	-1.335^{***}
(Higher = more literate)	(0.258)
Accurate Priors	0.675^{***}
(Higher = more accurate)	(0.236)
Political Knowledge	-0.069
(Higher = more knowledge)	(0.047)
WhatsApp Use Frequency	0.181^{***}
(Higher = more usage)	(0.049)
Trust in WhatsApp	-0.014
(Higher = more trust)	(0.059)
Education	0.045^{**}
	(0.019)
Male	0.295^{*}
	(0.167)
Hindu	-0.340^{**}
	(0.148)
Constant	10.400^{***}
	(0.443)
Observations	1,158
R^2	0.071
Adjusted R^2	0.063
Residual Std. Error	1.542 (df = 1147)
F Statistic	8.725^{***} (df = 10; 1147)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table G.2: Effect of Age on Digital Literacy

	Dependent variable: Digital Literacy (Higher = more literate)
	(1)
Age (Continuous)	-0.002^{***}
	(0.001)
BJP Supporter	0.020
	(0.013)
Accurate Priors (Higher = more accurate)	-0.063^{**}
	(0.027)
Political Knowledge (Higher = more knowledge)	0.051^{***}
	(0.005)
WhatsApp Use Frequency (Higher = more usage)	0.006
	(0.006)
Trust in WhatsApp (Higher = more trust)	0.024^{***}
	(0.007)
Education	0.007^{***}
	(0.002)
Male	0.079^{***}
	(0.019)
Hindu	-0.028
	(0.017)
Constant	0.376^{***}
	(0.049)
Observations	1,158
R^2	0.179
Adjusted R^2	0.173
Residual Std. Error	0.177 (df = 1148)
F Statistic	27.809^{***} (df = 9; 1148)
Note:	*p<0.1; **p<0.05; ***p<0.01

I now consider whether the literacy intervention worked better depending on age or digital literacy. In Table G.3 I interact the treatment with age and digital literacy, and do not find an interaction effect.

Table G.3: Effect of Treatment x Age and Digital Literacy

	Dependent variable: Number of Stories Accurately Classified	
	(1)	(2)
Literacy Intervention	0.386	0.349
	(0.306)	(0.382)
Age (Continuous)	0.032***	
	(0.009)	
Literacy Intervention x Age	−0.015	
	(0.011)	
Digital Literacy (Higher = more literate)		−0.984**
		(0.394)
Literacy Intervention x Digital Literacy		−0.454
		(0.489)
Constant	10.797***	12.378***
	(0.257)	(0.306)
Observations	1,224	1,224
R^2	0.016	0.025
Adjusted R^2	0.014	0.022
Residual Std. Error (df = 1220)	1.592	1.585
F Statistic (df = 3; 1220)	6.673***	10.270***
Note:		*p<0.1; **p<0.05; ***p<0.01

H True Stories

The outcome measure for this study comprised of more false stories than true (rather than a 50-50 split between true and false stories). This was done to maximize reducing belief in as many false stories as possible. However, several steps were taken to ensure that the imbalance of true vs. false stories did not affect the efficacy of the treatment. Before measuring the outcomes, respondents were told that some of the stories were false and some true, likely reducing the urge to default to the stories being false. Further, with the comprehension check, respondents were not only asked whether stories were true or false but were also asked how they identified the veracity of these stories. Importantly, a majority of respondents in the treatment groups said that their responses were motivated by enumerators teaching them about these stories during the household visit, rather than having learnt about the stories on the news or through a friend. Further, enumerators were instructed for this question to not read out response options aloud, but to allow respondents to organically speak about their views on the false stories in a way that minimized the ability of respondents to provide socially desirable answers. Thus, much care was taken in the experiment to ensure that the skew towards false stories would not impact respondents' answers.

I now analyze whether the treatment worked for the two true stories alone. Results are in Table H.1. I find that the perceptions of veracity of these stories did not depend on the treatment. However, respondents accurately classified a high proportion of the true stories, 76% and 95% respectively.

Table H.1: Identification of True Stories

	Dependent variable: Accurate Identification	
	1st True Story	2nd True Story
Literacy Intervention	−0.009	0.005
	(0.026)	(0.013)
Constant	0.776***	0.951***
	(0.021)	(0.010)
Observations	1,224	1,224
R^2	0.0001	0.0001
Adjusted R^2	−0.001	−0.001
Residual Std. Error (df = 1222)	0.421	0.209
F Statistic (df = 1; 1222)	0.134	0.171
Note:		*p<0.1; **p<0.05; ***p<0.01

I Predicted Story Identification

Here I visualize the interaction heterogeneous effect from partisan identity. I graph the predicted values from the interaction model in the bar plots below. In Figure I.1 I plot the predicted number of stories identified among pro-BJP stories; in Figure I.2 I do the same for anti-BJP stories.

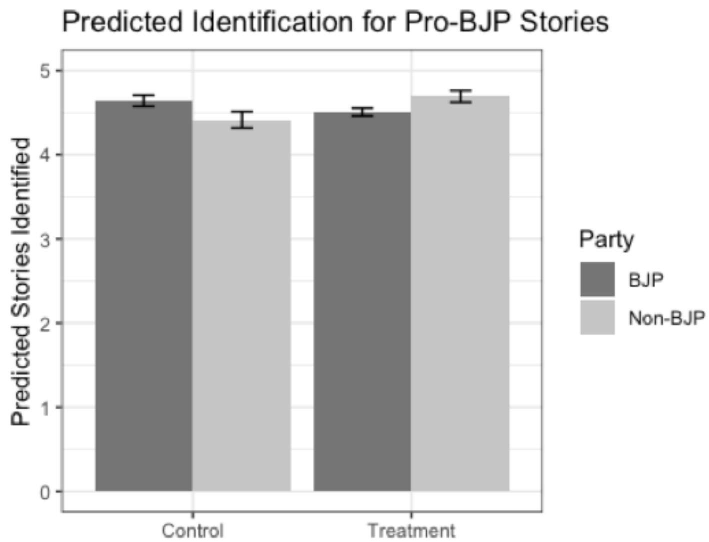


Figure I.1: Predicted Identification of Pro-BJP Stories

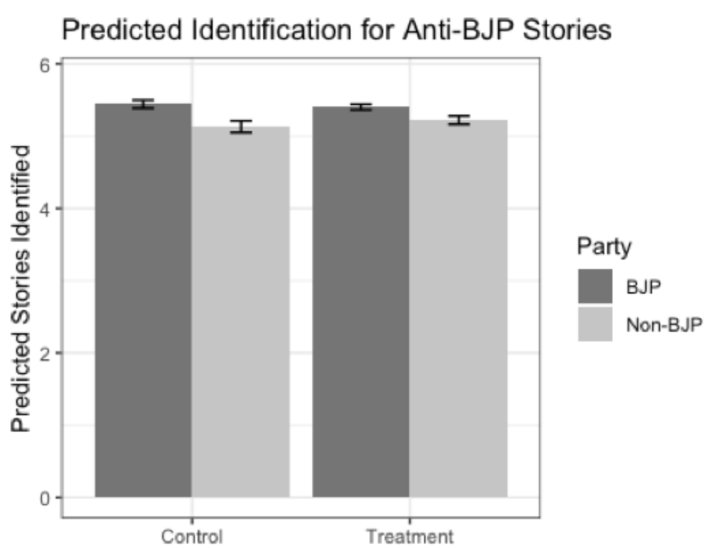


Figure I.2: Predicted Identification of Anti-BJP Stories

J BJP Partisanship

I explore in this section the finding that while BJP partisans are more likely to identify rumors correctly, pro-BJP rumors are also more likely to be believed. To better understand this result, I demons rate that BJP supporters' better identification is driven by their ability to identify anti-BJP stories as fake. In Table J.1 below, I list the rate of identification (average number of stories correctly identified) for BJP and non-BJP supporters for the two categories of stories that make up the DV, pro-BJP and anti-BJP rumors. The results from a two-sample t-test demonstrate that for pro-BJP stories, the difference in means is not different from 0. But for anti-BJP stories, the true difference in means is significantly greater than 0, with BJP respondents correctly identifying stories at a higher rate. Thus the finding that BJP respondents identify stories at a higher rate is driven only by their identification of anti-BJP messages as false.

Table J.1: T-test for Discernment by Party ID

	Mean for BJP	**Mean for non-BJP**	**T**	**p-value**
Pro-BJP Stories	4.55	4.59	0.69	0.488
Anti-BJP Stories	5.41	5.19	-3.63	0.000

Parsing this result further, I now examine whether salience of stories is linked with party-congruent beliefs. I examine whether there is a partisan divide in highly believed stories. To do so, I limit the dependent variable to the top two most believed stories (belief close to 50% of the sample); both stories are pro-BJP in slant. I find that for these stories, BJP supporters do significantly worse at discernment, i.e., they are significantly more likely to think that these false stories are true. I report this result in Table J.2.

Table J.2: Discernment of Top Two Believed Stories

	Dependent variable: Number of Stories Accurately Classified
	(1)
BJP Supporter	-0.130^{***}
	(0.045)
Constant	1.171^{***}
	(0.037)
Observations	1,224
R^2	0.007
Adjusted R^2	0.006
Residual Std. Error	0.727 (df = 1222)
F Statistic	8.458^{***} (df = 1; 1222)
Note:	*p<0.1; **p<0.05; ***p<0.01