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

Sumitra Badrinathan†
University of Pennsylvania

January 22, 2020

Abstract

Misinformation is a growing concern across the world, particularly in developing countries, where false claims can distort public opinion, undermine faith in democracy, and endanger public health. Despite its real-world import, little is known about how to combat this problem outside of the U.S. This study tests the efficacy of a persuasive literacy and fact-checking intervention on respondents' ability to identify fake news during the 2019 elections in India, using a field experiment in Bihar (N=1200). Treatment group respondents are provided with in-person media literacy training where enumerators demonstrate tools to identify fake news: reverse Google searching an image and navigating a fact-checking website. After an hour-long media literacy intervention I find no changes in respondent ability to identify fake news. However, results also demonstrate a partisan backfire effect: treated BJP respondents become less able to identify pro-attitudinal fake news. These findings point both to the stickiness of misinformation in India as well as the consolidation of partisan identity. This is the first paper of its kind to explore the impact misinformation has on opinions in India, making it an important contribution to political communication in a country where fake news has led to violence and murder.

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

*This study was preregistered with Evidence in Governance and Policy (20190916AA) and received IRB approval from the University of Pennsylvania (832833). The author thanks Devesh Kapur, Guy Grossman, Neelanjan Sircar, Marc Meredith, Matt Levendusky, Simon Chauchard, Yphtach Lelkes, Simone Dietrich and Devra Moehler. Pranav Chaudhary and the Sunai Consultancy team 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 fund.

†sumitra@sas.upenn.edu

1 Introduction

Images of mutilated bodies and lifeless children proliferated across WhatsApp chats in northern India in 2018, supposedly resulting from an organ trade racket by kidnappers. In response to these messages, a young man mistaken to be a kidnapper was mobbed and brutally beaten by village locals in Meerut, Haryana.[1] Similar incidents occurred in sixty-one separate villages in a 6-month span in 2018, and in 33 of those cases village mobs lynched innocent men to death on mere suspicion of them being child kidnappers.[2] The photos and videos that provoked these brutal murders, although not fabricated, constitute misinformation: the children in the pictures had actually been killed in Ghouta, Syria, five years ago rather than as part of the organ trade in India, like the message recipients believed.[3] Though media outlets and local law enforcement debunked the child kidnapping videos after the first lynchings, the attacks continued nonetheless. Angry mobs committed violent murders in neighboring towns as the messages went viral from village to village, demonstrating that fake news can do more than sway opinion and elections: in India it is a public health crisis.

What, if any, effective tools exist to combat this public health crisis? This paper explores the effectiveness of a field intervention in countering misinformation in India. After the 2016 American election, a slew of academic articles tested and measured fact-checking interventions to address fake news, including providing warning labels on articles, attaching disputed tags to misinformation, and providing corrective information beside fake claims (Nyhan et al 2017, Clayton et al 2019, Bode and Vraga 2015, Pennycook et al 2017). The authors of these studies share the common belief that fake news is an existential threat and that possessing incorrect information comes at a cost to the health of democracy. However, given that the impacts of fake news in India go as far as violence and murders, I argue that we need to treat the misinformation crisis as a public health problem, not very different from the way in which we combat infectious diseases. While these fact-checking interventions have been shown to be effective at treating misinformation in developed countries, I argue such technical fixes are unlikely to be sufficient to deal with the scope of fake news in India The last few years has seen the growth of several fact-checking platforms in India, but to truly fight the misinformation battle, we need more than mere technical fixes: mass educative and persuasive programmes are necessary to deal with the menace of misinformation. Technical fixes might work in the short run, but in countries with low literacy rates and increasingly advanced internet connectivity, the most promising long-term solutions involve costly, mass education programs about the negative effects of misinformation and tools to combat it, not unlike like the persistent mass

[1] Ahmed, Qazi Faraz. (2019). The Rumour That Killed: Incidences of Mob Violence over Child-lifting Hoax on Rise in Uttar Pradesh. *News 18*, August 28. https://www.news18.com/news/india/the-rumour-that-killed-incidences-of-mob-violence-over-child-lifting-hoax-on-rise-in-uttar-pradesh-2286987.html

[2] Saldanha, A., Rajput, P., Hazare, A. (2019). Child-Lifting Rumours: 33 Killed In 69 Mob Attacks Since Jan 2017. *IndiaSpend*, July 9. https://www.indiaspend.com/child-lifting-rumours-33-killed-in-69-mob-attacks-since-jan-2017-before-that-only-1-attack-in-2012-2012/

[3] McLaughlin, Timothy (2018). How WhatsApp Fuels Fake News and Violence in India. *Wired Magazine*, December 12. https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/

education used to curb, and eventually eradicate, polio in India back in 2012.[4]

The specific research question asked in this paper is whether a strong persuasive and digital literacy intervention is effective in combating fake news in India. To answer this question, I conducted a field experiment with 1200 respondents in Bihar, India during the 2019 general elections, when misinformation is arguably at its peak. The experiment involves an education, literacy training and fact-checking phase, followed by a measurement phase. In an hour-long intervention, treatment group respondents were shown demonstrations of two tools to verify information: reverse-google searching images, and navigating a fact-checking website. They also received several tips to prevent falling for fake news. Treatment group respondents were also given a flyer with these tips, along with examples of four political fake stories with their corrections. After a two week period during which respondents voted in the 2019 general election, I revisited respondent households to measure their ability to identify fake news. Respondents were given a series of news stories that varied in salience, political slant, and veracity. For each story, I measured respondent ability to identify the story as fake as well as confidence in that ability. Secondary dependent variables measures include self-perceived ability to spot misinformation, trust in WhatsApp, and a clean measure of voter turnout.

My experiment shows that an hour-long, strong persuasive and educative treatment is not sufficient to help respondents combat fake news. Importantly the average treatment effect was not significantly distinguishable from zero even though the experiment occurred during the election (when the most fake news is pedaled) and in Bihar (the state with the lowest literacy rate in India). Finding that a bundled, hour-long and costly intervention does not move people's prior attitudes on fake news is testimony to the permeating, indestructible and detrimental effects of misinformation. My findings challenge conventional wisdom in American politics that subtle priming treatments such as disputed tags can reduce fake news consumption. These findings also confirm qualitative evidence about the unique nature of social media in developing states that are new to the internet lending them particularly rife and vulnerable to misinformation.

While there is no evidence of a non-zero average treatment effect, there are significant treatment effects among some subgroups. BJP partisans in my sample possess a highly consolidated partisan identity: they are less likely to be able to identify news as fake when they receive the intervention, but only when the false news paints their party in a good light. These partisans do not demonstrate similar misinformation identification capacities for anti-BJP fake news. Theoretically, this illustration is akin to a partisan backfiring effect: the treatment appears to make BJP partisans less likely to identify pro-partisan stories as fake; this effect does not obtain for non-BJP partisans at all, or for BJP partisans and pro-BJP fake stories. This is consistent with findings in the American politics literature demonstrating that partisans cheerlead for their party and are likely to respond expressively to emotionally loaded questions (Gerber and Huber 2009; Prior, Sood, and Khanna 2015). It also challenges the contention that Indians lack consolidated, strong partisan identities

[4]UNICEF Press Release (2011). Amitabh Bachchan launches new Polio Communication Campaign. *UNICEF*, December 19. http://unicef.in/PressReleases/69/Amitabh-Bachchan-launches-new-Polio-Communication-Campaign

(Chibber and Verma, 2018). I demonstrate that party identity in India is more polarized than previously thought, particularly with BJP partisans.

This is the first analysis of the effect of educative and persuasive interventions on misinformation in India. To the extent that consuming misinformation imposes costs and given that a strong, in-person treatment does not bring desired results, this study hopes to spark a research agenda on the ways to create an informed Indian citizenry armed with tools to fight misinformation as the commercial and partisan incentives to spread fake news increase in the near future. I also seek to contribute to the empirical study of partisan identity in India, countering the conventional wisdom of party identities being unconsolidated and fluctuating. In this arena, I present evidence from the first large-scale field education intervention to repel fake news and misinformation in India.

This paper begins in the next section by discussing the extent of misinformation and the tools employed thus far to combat it. It then reviews relevant features of the Indian context that make it a unique hotspot for misinformation, the experimental design, and the media literacy intervention for this study. I next present analysis for the main effects of the treatment on the key dependent variable as well as heterogeneous effects of partisan identity. I conclude with implications of these results and future directions for the research on misinformation.

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

Following Alcott and Gentzkow (2017), I define "fake news" as stories that are intentionally and verifiably false and could mislead readers. This includes both false news and fabricated stories presented as legitimate news, as well as hyper-partisan news, and biased, misleading interpretations of actual news. In this paper, I focus only on the subset of these stories that are political in nature, with special attention given to those stories that became viral in the months preceding the Indian general election of 2019.

The focus on fake news and misinformation in the 2016 American election has provoked considerable attention in mainstream media as well as academic literature, awakening a "sleeping giant" in political communication research (Weeks, 2019). Attempts to study where misinformation comes from, its prevalence, why it is believed, and its effects on public opinion, among other things, have led to a rapid and expansive outbreak of scholarly attention on fake news.

While fake news is now a global phenomenon, this literature is grounded in the American context, with a particular focus on the 2016 election. The literature finds that misperceptions are widespread: Flynn (2016) posits that over 20% of Americans confidently believe misinformation about political issues. Examples of this include widespread beliefs that certain vaccines can cause autism in healthy children, or obstinate reliance in the false information that President Obama was not born in the US (Freed et al 2010; Nyhan 2012). Holding incorrect beliefs has been shown to affect political behavior when those beliefs are associated with political choices or actions. Hochschild and Einstein (2015) theorize that the combination of holding incorrect knowledge as well as actively participating in politics is the most concerning: if people in this category held accurate beliefs, their opinions and behaviors might be different from what we currently observe.

Starting from this observation is the assumption that misperceptions impose private and social costs to democracy by disallowing citizens to infer the true state of the world (Alcott and Gentzkow, 2017). Given its normatively troubling consequences,, a vast research agenda has tested and measured interventions to combat fake news and reduce the consumption of misinformation. Such studies typically rely on providing corrections, warnings, or fact-checking treatments to respondents and consequently measuring their 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, 2016); it used this approach for a year before switching to providing fact checks underneath suspect stories (Smith et al, 2017). The prevalence for piloting such "quick fixes" to the misinformation problem has since exploded: Chan et al 2017 find that explicit warnings can reduce the effects of misinformation; Pennycook et al. (2017) test and find that disputed tags alongside veracity (accurate / not accurate) 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.

Overall, however, this research has been met with mixed success: fact-checking and warning treatments are only effective when misinformation is not salient, when priors are not strong, and when outcomes are measured immediately after intervention. Corrective information fails to change beliefs when the individuals hold strong priors and when the information being corrected is salient, leading to the most significant misperceptions being stable and persistent over time (Nyhan 2012; Flynn, Nyhan and Reifler 2016).

Such persistence stems from respondent's motivated reasoning (Flynn et al., 2017). Human reasoning is pulled between the Scylla of accuracy and the Charybdis of confirmation: humans want to come to the "right" answer, but the architecture of our minds biased us toward information that reinforces our prior beliefs (Kunda, 1990). As a result, we tend to seek out information that reinforces our preferences (confirmation bias), counter-argue information that contradicts preferences (disconfirmation bias), and view pro-attitudinal information as more convincing than counter-attitudinal information (Taber and Lodge, 2006). Citizens face a tradeoff between a private incentive to consume unbiased news and a psychological utility from confirmatory news, resulting in a diminished effect of corrective interventions (Gentzkow, Shapiro and Stone, 2016). In particular, political contexts – where partisan identities heighten our tendency to seek out like-minded information – accelerate this underlying tendency. Because of motivated reasoning, misperceptions persist over time when they reinforce a respondents' views (Nyhan, 2012), and exacerbate the effects of false information even after it has been debunked (Bullock 2007; Thorson 2015a).

Not surprisingly, then, despite the staggering number and comprehensive set of variations in fact-checking and warning treatments, such studies have had little success in combating misinformation over time. Further, even setting aside the psychological challenge, these studies are almost all lab and survey experimental, and hence we know little about their real-world ecological validity. Empirical findings from studies on media effects that use lab settings may not be reliable indicators of the effects observed in the real world (Jerit, Barabas and Clifford, 2013). Moreover, the success of lab and survey experiments has policy implications limited to populations who are frequently online and use platforms such as Facebook and MTurk. This does not describe the vast majority of populations in developing countries, who are both less likely to be avid internet users and more likely to fall prey to misinformation. Reacting to these challenges, review articles express hopefulness about the potential of finding stronger treatment effects if the corrective interventions were undertaken in a field setting (for example, Flynn, Nyhan and Reifler, 2017).

The body of academic research on measures to correct misinformation in the United States constitutes an important contribution to scholarship but simultaneously highlights the dearth of high-quality field interventions to combat misinformation. In doing so, it calls for four key modifications that misinformation studies should undertake: changing the methodological approach from a lab or survey setting to a field intervention, moving beyond subtle technical fixes to providing strong, educative treatments, measuring the effect of treatments over time and in highly salient fake news settings, and finally, testing interventions in contexts of low literacy and recent media penetration. In the following sections, I will discuss how this paper develops an intervention addressing these challenges to combating misinformation in India.

3 Misinformation in India

Misinformation is not uniquely an American problem. Studies show that Bolsonaro benefited from fake stories online, misleading information was used to incite violence against the Rohingya in Myanmar, and minorities are frequently targeted in Sri Lanka because of online rumors.[5] Though false and misleading stories are worrisome for advanced democracies, they portend a stronger and more immediate crisis for developing and emerging economies – research demonstrates that fake news "played an important role" in elections in at least 17 other states apart from the US, most of them developing countries (Freedom House, 2018).

More importantly, important differences between the U.S. and Indian contexts mean that the theoretical framework used in the US may be insufficient for the politics of the Indian subcontinent. The predominant model of misinformation consumption in the literature comes from Gentzkow, Shapiro and Stone (2015). They posit that consumers have an incentive to prefer confirmatory stories rather than the truth because of the psychological utility they get such stories, thereby producing an equilibrium where news outlets are incentivized to report on stories in a biased way (consistent with a motivated reasoning account, as discussed above). In the model, consumers must make a choice between deriving psychological utility from ideologically-consistent news, or receiving utility from knowing the true state of the world. In order to make this tradeoff, the model implicitly assumes that consumers are aware that they have a choice to make and can choose their actions with complete knowledge of the tradeoff they are having to engage in. Though this model may comprehensively explain the state of the developed world, it falls short in explaining

[5] Sanchez, Conor. (2019). Misinformation is a Threat to Democracy in the Developing World. *Council on Foreign Relations*, January 29. https://www.cfr.org/blog/misinformation-threat-democracy-developing-world

contexts where consumers may not be fully aware of the existence of false news and misinformation, thereby taking away the tradeoff between motivated reasoning and accuracy concerns. Theoretical viewpoints that support such models may conclude that consumers who fall for misinformation are not merely biased, but simply lazy (Pennycook and Rand, 2017). But the conscious choice forgo true information and the "laziness" associated with consuming false information may both not characterize media markets where consumers are not aware about the existence of fake news to begin with. This lack of awareness could stem from several factors that I describe below.

The menace of fake news is in India is compounded by the presence of four factors that make the country a uniquely vulnerable hotspot in the global misinformation crisis. The first is rates of literacy and education. India has had a traditionally lower literacy rate, both in comparison with other countries in South Asia as well as relative to developing countries across the world where fake news has been shown to affect elections and public opinion. In 2015, India's literacy rate was estimated to be around 72%, lower than Kenya (78%), Sri Lanka (93%), Myanmar (93%), and Brazil (93%).[6] Further, India also ranks relatively low in its share of population with no formal education: about 30% in 2015, relative to 14 percent or less in Myanmar, Brazil, Sri Lanka. To add to this, the sample site for this study – the state of Bihar in India – has historically had one of the lowest literacy rates within the country (currently at 63%). These low education and literacy rates aggravate the misinformation crisis in India, given that studies demonstrate that people with higher education have more accurate beliefs about the news (Alcott and Gentzkow 2017; Nyhan and Reifler 2017). Hence we should expect a higher vulnerability on average to misinformation and misperceptions among populations with lower literacy and lower education.[7]

Second, the expansion of digital economy in India has meant that more users are now connected to the internet in the country than ever before. India is now one of the largest and fastest-growing markets for digital consumers, with 560 million internet subscribers in 2018, second only to China. An average mobile data user in India currently consumes more than 8 GB of data per month, which exceeds the average in more digitally advanced countries such as China and South Korea. Estimates show that the number of users connected to the internet could exceed 800 million by 2023, 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). Recent reports demonstrate that the penetration of the internet in rural India increased from merely 9% in 2015 to 25% in 2018, connecting over half a billion people in the country to the internet, with rural India alone registering over 250 million users. Bolstered by some of the world's cheapest data and bandwidth plans along with systematic and persistent governmental efforts to increase rural connectivity, the state of Bihar – the sampling site for this study – alone saw an internet connectivity growth of over 35% in 2018, the highest in the country.[8]. India is digitizing faster than most mature

[6] Roser, M. and Ortiz-Ospina, E. (2018). Literacy. *Our World in Data*. https://ourworldindata.org/literacy

[7] The role of education in affecting correct beliefs is debated – research demonstrates that it can increase ability to tell fake news from reality, but motivated reasoning gives respondents better tools to against counter-attitudinal beliefs (Nyhan and Reifler 2017).

[8] Press Trust of India (2019). Internet users in India to reach 627 million in 2019: Report. *The Economic Times*, March 06. https://economictimes.indiatimes.com/tech/internet/internet-users-in-india-to-reach-627-million-in-2019-

and emerging economies.

The emerging literature on Internet Communication Technologies (ICT) and mobile technology in developing settings (Brown, Campbell, & Ling, 2011; Donner & Walton, 2013; Gitau, Marsden, Donner, 2010) finds several avenues through which mobile devices can improve digital inclusion and learning, economic development, and quality of life. Paradoxically, this leap in development might also mean that the novelty and unfamiliarity of the medium coupled with the fascination it inspires makes users more vulnerable to the information they receive online. The obstacles to information-seeking on mobile devices might paint a cautionary tale. Research demonstrates that obtaining accurate information on mobile devices is costlier than other mediums (Donner and Walton, 2013) and that mobile-driven information attenuates attention paid to news (Dunaway et al, 2018). 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), hence the problem of misinformation in India is further compounded.

Take, for example, the case of Geeta, who lives in the suburbs of Ara in Bihar in a 11x14 feet chawl along with her 3 sons. Cheaper mobile data plans have finally allowed Geeta and her sons to buy a smartphone, which they all share. Geeta is a housewife who only studied until grade 5, but says she enjoys the WhatsApp photos and videos she receives in group chats and is adept at using the phone's text-to-voice to respond to them. I ask her whether she thinks the photos and videos she receives through the app are all factually correct. She says:

"This object [her Redmi phone] is only the size of my palm but is powerful enough to light up my home so I do not even have to turn on the bulb. Previously we would have to walk 1.5 kilometers to the corner shop that has a tv to learn about politics and news. Now when this tiny device that fits in my hand 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][9]

Geeta's example demonstrates that the novelty of digital media increases might increase vulnerability to all kinds of information and poses a greater challenge to combating misinformation. Recent studies on emerging economies underscore this point by demonstrating that in countries such as India, many people use the internet without realizing it. Substantial numbers of people in India say they do not use the internet but also indicate having a phone that connects to the internet; or that they are not connected to the internet but use social media (Silver and Smith, 2019). The presence of such "unconscious" internet users who appear to be using the internet without understanding that they are going online demonstrates that in developing countries such as India, people may be unaware of what the internet is in a variety of ways. For users such as Geeta, the smartphone constitutes a novel sensory experience that is both more likely to entice users while simultaneously attenuating attention to information received online.

Third, a key factor differentiating misinformation in India from other contexts is the medium through which most of it is sent and received. Unlike the US, fake news spreads mainly on chat-

report/articleshow/68288868.cms

[9]Interview with Geeta, 27 March 2019. All names have been changed; all quotes translated from Hindi

based apps in India, the most common of which is WhatsApp. According to WhatsApp itself, Indians users forward more messages, photos and videos than in any country in the world. The result is that users on WhatsApp are subjected to a seemingly unending barrage of information, ranging from purely factual (cricket scores), to largely harmless information catering to credulous people (horoscopes), to negative stereotypes of social groups, and finally to malevolent hate-forming messages. India is WhatsApp's biggest market in the world, with over 400 million users connected to the app, reaffirming its popularity and gigantic reach.[10] However a more important reason contributing to the app's popularity is also at the heart of the fake news problem: WhatsApp messages are private and protected by encryption. This means that no one, including the app developers and owners themselves, have access to see, read, filter, and analyze text messages. The consequence of this feature is that tracing the source or the extent of spread of a message in a network is close to impossible, making WhatsApp akin to a "black hole" of fake news.

Finally, the content of fake news in India is itself unique. While misinformation in the United States is spread mostly over social media sites such as Facebook and takes the form of a news article on a website, most new that goes viral on WhatsApp in India largely constitutes photo-shopped images and manufactured videos, making it both harder to counter through traditional fact-checking mechanisms and easier to believe (Sinha, 2019). Misinformation in graphical and visual form has been demonstrated to have increased salience and retain respondent attention better (Flynn, Nyhan and Reifler 2015), thus countering fake videos and photos is a harder task than countering plain text. The image below is an example of a fake story that circulated over WhatsApp claiming that cow urine cures cancer. The viral news forward is the image itself, in which an apparent patent is demonstrated as proof of linkages between consuming cow urine and curing cancer. The text at the bottom of the image paints a picture of how such viral rumors, even those about health, are often political in nature. The text associated with fake and misleading photos and videos in India often does not emanate from a particular news source, but is usually simply a forwarded message on a WhatsApp group that one received from a friend or a family member. Hence traditional fact-checking measures that prescribe checking the link of a story to find red flags will not work in this context because the rumor is the image or video itself. Thus the nature of fake news in India is different and unique and hence countering it requires measures that are stronger than fact-checking, that are tailored specifically to countering misleading photos and videos in the WhatsApp world.

New access to internet coupled with low education rates and the privacy and encryption of WhatsApp together serve to magnify the already severe misinformation crisis in India. These features of the Indian context suggest the need to take a step back in theorizing misinformation markets and begin from the starting point where the consumer is *not* assumed to understand the costs and benefits of choices involving tools and technology that are new and unfamiliar. Solutions to the fake news problem need to hence be updated to reflect the needs of a citizenry that is less

[10] Singh, Manish (2019). WhatsApp reaches 400 million users in India, its biggest market. *Tech Crunch*, July 26. https://techcrunch.com/2019/07/26/whatsapp-india-users-400-million/

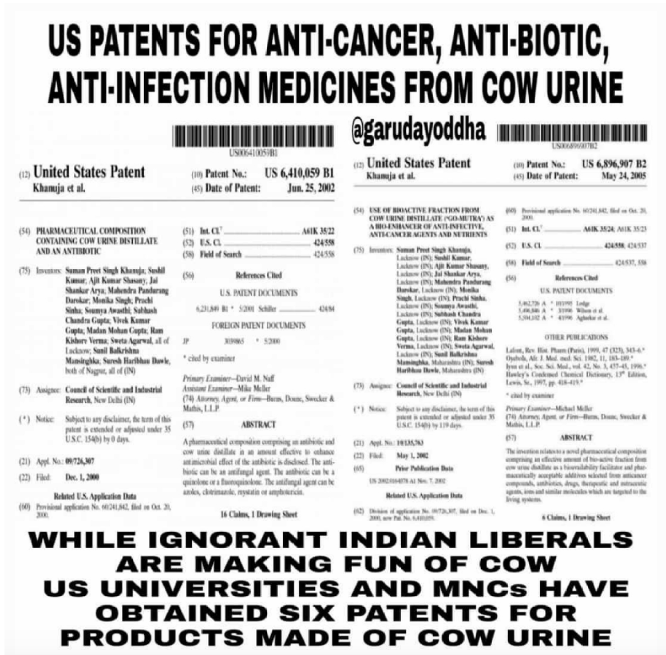


Figure 1: "Cow Urine Cures Cancer" Viral WhatsApp Rumor

educated, far newer to the internet, and where primary modes of information consumption are group chat applications. Hence, traditional means of using fact-checking and technical fixes in online and survey experiments might fall short in dealing with the misinformation menace in India.

To deal with the menace of fake news, the company WhatsApp launched a campaign in 2018 including newspaper and television ads focusing on user-driven fake news identification strategies. The newspaper advertisements included tips to spot fake forwards on WhatsApp,[11] and the television ads included visuals exhorting users to forward only messages they are convinced are true.[12] The newspaper ads, however, appeared only in English and are thus unlikely to have been seen by the vast majority of WhatsApp users in India. Further, they include tips such as "check elsewhere to see if a story is true" and appeals to users to question information that they receive via the app. While useful in settings where users know how to verify news online and understand what it means to "question" the veracity of a story, the practicality of this solution is limited in contexts of lower digital literacy and unfamiliarity with the internet. None of the forty-five respondents interviewed in focus group settings for a pilot for this project said they had actually seen the WhatsApp newspaper ads or tv commercials. When we explained some of the tips included in the WhatsApp ads, several respondents said they understood from the ad that news can be misleading, but are still unsure about how to verify news. One respondent admitted that her only source of news was WhatsApp groups, and hence she wouldn't know where else to verify the messages she receives.

[11] Firstpost Tech2 News Staff (2018). WhatsApp Rolls Out Newspaper Ads With Tips For Users To Fight False Information. *Firstpost*, July 10.https://www.firstpost.com/tech/news-analysis/whatsapp-rolls-out-newspaper-ads-with-tips-for-users-to-fight-false-information-4702911.html

[12] Thakkar, A. (2018). WhatsApp is now betting on Indian primetime television to fight fake news. *Quartz India*, December 04. https://qz.com/india/1482656/whatsapp-launches-tv-ad-campaign-to-fight-fake-news-in-india/

Another highlighted that the sentence "question information that upsets you" was vague and did not truly indicate what to do about such information. These focus group discussions underscored that the user-driven WhatsApp solutions may not encompass the vast majority of users for whom the internet is novel, and that to fully apply the techniques included in the advertisements, users would need a fuller primer to the tangible ways in which they can verify information.

This paper is an attempt at bridging the gap between solutions to misinformation in the western world and their applicability to emerging digital economies such as India. Given the recent advances in internet penetration in India coupled with platforms where messaging is encrypted, I propose the use of persuasive, in-person treatments with special focus on educative tools that can combat fake news in the Indian context. This setup is a natural extension of the existing solutions proposed by WhatsApp, but allows for learning in a one-on-one session and provides users with tangible methods that they can use to ascertain the veracity of information. Mass education programs to eradicate public health problems in India have been shown to have successful outcomes in fostering long-term desired outcomes – India's successful track record of such public education efforts include family planning in the 50s, anti-polio campaigns, and more recently, a government policy that convinced millions of people to voluntarily give up their gas subsidies.[13] In the next section I describe the intervention and design of this study to tackle fake news with a strong, educative, in-person treatment.

4 Research Design

Although persuasive in-person interventions are widely applied in peace-building, conflict and public health settings, to my knowledge, the method has never been directly used to fight misinformation in an empirically rigorous field experiment.

4.1 Context and Research Setting

Social media data from the months before the 2016 American election shows that the most popular fake news stories were more widely shared on Facebook relative to mainstream stories, and that the most discussed fake news stories were political in nature, favoring either Donald Trump or Hillary Clinton (Silverman, 2016). This finding is not very different from the Indian context, where fake news is purported to be most salient during and before elections, with the goal to spread misinformation that could affect voting behavior (Sinha et al, 2019). The context of this study is that of the consequential 2019 general election in India, where both mass polarization and fake news rose to the extent that observers called it a "WhatsApp Election".[14]

The election context of fake news dissemination in India is unique – unlike the United States

[13]Daily Excelsior (2016). PM thanks 1 crore people for giving up LPG subsidy to help poor. April 23. https://www.dailyexcelsior.com/pm-thanks-1-crore-people-giving-lpg-subsidy-help-poor/

[14]Bengani, Priyanjana. (2019). India had its first 'WhatsApp election.' We have a million messages from it. *Columbia Journalism Review*, October 16. https://www.cjr.org/tow$_c$*enter/india − whatsapp − analysis − election − security.php*

where the focus has been on foreign-backed misinformation campaigns, the fake news circulating in India aimed at affecting elections and public opinion is not manufactured abroad. Most of India's political fake news circulated over WhatsApp is produced and spread by political campaigns themselves (Singh, 2018). The 2019 contest was a reelection bid for Narendra Modi, leader of India's Hindu nationalist Bharatiya Janta Party (BJP). India is a parliamentary system, but Narendra Modi's style of politics makes it akin to presidential elections with a high level of polarization, where not unlike Donald Trump, he inspires either fervent loyalty or deep distrust.[15] To add to this, qualitative evidence suggests that parties like the BJP hired citizens specifically to promote the party on WhatsApp chats (Singh, 2018), with BJP president Amit Malviya stating on record that state units were asked to compile lists of smart-phone carrying voters such that they could be added to WhatsApp groups, one for each polling booth.[16] Often, these groups morph into havens for misinformation, sometimes pedaled by party workers themselves (Singh, 2018).[17]

A final differentiating factor between the American and the Indian context is the role of motivated reasoning. Research from the US demonstrates that the most common sources of directional motivated reasoning are partisanship and strength of party identity (Nyhan and Reifler, 2010). But it is unclear whether this theory and findings carry forward to the Indian context. Research in the Indian context describes India as a traditionally and historically non-ideological system, analysing how politics in India is conventionally viewed as chaotic and volatile (CHibber and Verma, 2008; Chandra, 2004, Wilkinson and Kitschelt, 2007). Indian politicians have made and un-made coalitions with little regard to the partners with whom they have aligned. Institutions over time have been subjugated to individual interests rather than collective party interests. Given this observation, one might argue that the Indian case presents the opposite of the Michigan School's American Voter model, in that parties need not singularly dictate political attitudes. This casts some doubt on whether partisan motivated reasoning operates in the same way in India that it does in the American context. However the timing of this study during the election, where qualitative and anecdotal evidence suggests that party politics is polarizing, lends itself to potentially interesting party effects on the treatment.

All this suggests implementing a costly, in-person intervention to fight fake news at the time when it is most salient and pervasive – as in this study – is a hard test for the goal of reducing misinformation consumption. The sample of respondents in this study is drawn from the city of Gaya and its suburbs in the state of Bihar in India. Bihar has both the lowest literacy rate in the country as well as the highest rural penetration of mobile phones. The rationale behind choosing the city of Gaya as the base for sampling was to be able to roll out the intervention in a city that is urban yet agricultural, with a high degree of internet usage. Gaya is the second largest city in

[15] Masih, N. and Slater, J. (2019). U.S.-style polarization has arrived in India. Modi is at the heart of the divide. *The Washington Post*, May 20. https://www.washingtonpost.com/world/asia$_pacific/divided-families-and-tense-silences-us-style-polarization-arrives-in-india/2019/05/18/734bfdc6-5bb3-11e9-98d4-844088d135f2_story.html$

[16] Poonam, S. and Bansal, S. (2019). Misinformation Is Endangering India's Election. *The Atlantic*, April 01. https://www.theatlantic.com/international/archive/2019/04/india-misinformation-election-fake-news/586123/

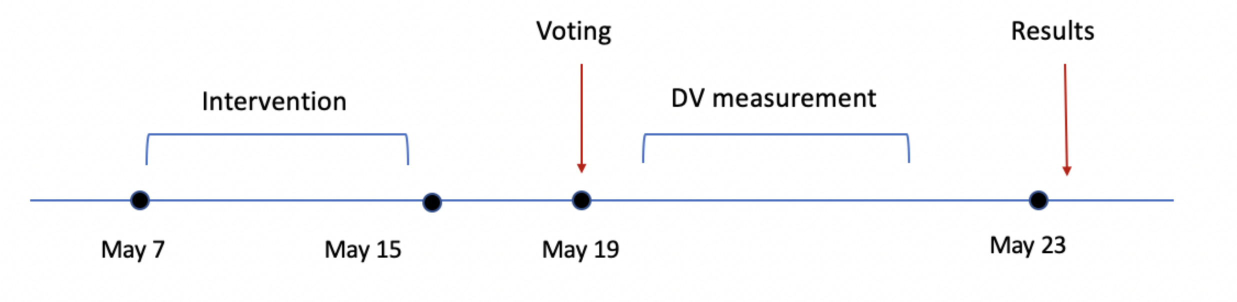
[17] The BJP is not the only party to do this, see link above

Bihar; it is rapidly urbanizing yet not as developed as the Tier 1 and Tier 2 metro-cities of the country (such as Mumbai and Delhi), making it relatively more descriptive of the average Indian city.

4.2 Experimental Timeframe

The baseline survey and experimental intervention for this study was rolled out in the first two weeks of May 2019 in the city of Gaya and its suburbs. Approximately two weeks after the intervention, we revisited the same respondents to conduct an endline survey and measure the outcomes of interest. Figure 2 summarizes the timeline for this study. The study took multiple steps in survey design and implementation to minimize exogenous shocks from election results. Notably, the study timeline ensured that though respondents voted in the general election *after* the intervention, making voter turnout post-treatment, the endline survey to measure outcomes was conducted *before* election votes were counted and results were announced. This timeline had the double advantage of ensuring that the DV 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.

Figure 2: Experimental Timeline



4.3 Survey and Sampling Design

The goal of this study is to test whether educative interventions are successful at helping WhatsApp users identify fake news. To maximize WhatsApp users who are familiar with the internet in the sample, I restricted the sample based on 3 variables. First, respondents were required to own their own cellphone (i.e. a personal phone for themselves and not a shared household phone). Second, the phone was required to have working internet for at least the last 6 months, and third, WhatsApp was required to be downloaded on the phone. These restrictions were put into place to ensure that respondents in the study were likely to be exposed to political fake news over WhatsApp in the months leading up to the election. The primary sampling unit, the city of Gaya and its suburbs, consists of several 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). These 85 polling booths served 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 that elections were being conducted. Surveying households within the each chosen polling booth area involved choosing a starting point and then proceeding along a path, selecting every k^{th} household. I follow 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 can be a street corner, a school, or a water source, and field enumerators were instructed to randomly rotate the choice of such landmarks. From the landmark starting point, the field enumerator walks in a designated direction away from the landmark and selects the tenth household for the survey, counting houses on both the left and the right. Once they leave their first interview they continue in the same direction, selecting the next household after another interval of ten. If the settlement comes to an end and there are no more houses, the field enumerator turns at right angles to the right and keeps walking, continuing to count until finding the tenth dwelling. Each field enumerator is assigned to only one polling booth, and hence the paths taken during each selection cross each household only once, providing a random and unbiased sample. Once a household is selected, a randomly chosen adult member (ages 18-60) of the household is selected to answer our survey questions once they qualify based on pre-conditions (smartphone access with Internet and WhatsApp for 6 months). The final sample comprised of 1224 respondents.

At the end of the baseline survey, enumerators collected addresses and mobile numbers of respondents 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. Enumerators interviewed respondents at their homes for both the baseline and endline surveys.

4.4 Experimental Design

The objective of this study was to design a persuasive, educative intervention that sought to reduce fake news consumption by empowering respondents with tools to identify fake news. To this end, the intervention targeted to treatment group respondents was, by design, a bundled treatment intended to have a strong effect on the outcome. It consisted of surveying a respondent in their home and undertaking the following activities in a 45-60 minute visit:

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

2. Demographics and pre-treatment covariates: The enumerator administered survey modules

designed to measure demographic details and pre-treatment covariates including digital literacy, political knowledge, media trust, information sources, and prior ability to discern fake news from real news.

3. Media literacy intervention: This was the key part of the study where enumerators demonstrated to respondents how to use two media literacy tools especially relevant to the Indian context to combat misinformation.

    - Reverse Google searching an image: Because a large amount of fake news in India comprises of misleading photographs and video clips, an easy way to learn whether a photo or video screenshot constitutes misinformation is to reverse Google search it. Feeding an image back into Google search can tell you that a photo of a missile that claimed to kill over 300 terrorists in Pakistan is actually a screen grab from a video game called Arma 2. It can also tell you that messages claiming that a parties are hacking elections that come with photos of voting machines being loaded into trucks are in reality photos from over two years ago and not proof of misconduct in the current election. The timestamp on an image or an article associated with an image can help dispel misleading information that the image constitutes current news, given that respondents can see the original source and time stamp on the image once it is fed back into Google. This is a uniquely useful tool for the Indian context given the nature of misinformation in India. Enumerators demonstrated two examples of this to respondents.
    - Navigating a fact-checking website: Enumerators also demonstrated how to use a fact-checking website (in Hindi) and showed respondents how most viral news stories get fact-checked on the site and appear in a ticker; they demonstrated how to use the search bar to find whether a particular story had been fact-checked by the site.

4. Fact-checking and tools flyer: The enumerator next handed out a flyer to respondents. The front of the flyer contained four recent political fake stories that went viral. The enumerator systematically debunked each story, explaining why the story was untrue, what the correct version was, and how we can determine the story is fake. The back side of the flyer contained six tips to reduce the spread of fake news. The enumerator read and explain each tip to respondents. Finally, the enumerator left the flyer behind with respondents, exhorting them to make use of it in the future when they try to determine whether a message they received over WhatsApp is fake or not.

5. Comprehension Check: Lastly the enumerator administered a short module of three questions that acted as a comprehension check aimed at measuring whether the treatment was effective in the short-term i.e. in immediately reducing the perceived accuracy of the fake stories that were debunked in the flyer.

Figure 3: T1: Pro-BJP Flyer (front and back)

The intervention had several features. First, it involved an in-person media literacy demonstration and then a fact-check, thus in some sense making it more akin to a persuasive intervention. Second, the intervention did not merely label stories as false or disputed. Enumerators explained in detail to respondents why each story was false and how they were debunked. This made the intervention more persuasive and credible to respondents. Third, a lot of fake news in India exists in the form of photoshopped images and false videos, hence I used graphics and visuals in order to increase the salience of news stories as well as retain respondents' attention. Rather than using traditional technical fixes or fact-checking methods, the media literacy tools used here are designed to combat the unique nature of fake news in the Indian context. All these factors made the treatment stronger (Flynn, Nyhan and Reifler, 2015). Further, the intervention provided respondents with concrete tips to verify information and identify fake news. Taken together, this bundled treatment was designed to provide the strongest possible intervention to combat fake news during the election.

Each household was revisited by the same enumerator for an end-line survey after two weeks. This end-line survey was intended to measure the effect of the intervention on respondents' ability to identify fake news.

4.4.1 Treatment Arms

Respondents were randomly divided into three groups, two treatment and one placebo control. The two treatment groups differed only in the flyer that they received. The table below summarizes the three groups:

Table 1: Experimental Treatments

Intervention	Goal
T1: Media Literacy + Pro-BJP flyer	Fact-check pro-BJP fake stories
T2: Media Literacy + Anti-BJP flyer	Fact-check anti-BJP fake stories
Control: Placebo flyer	Plastic pollution and tips to reduce plastic use

Note that the two treatment groups differed only in the stories that were fact-checked; both groups received the same media literacy training. The placebo control group received a flyer on plastic pollution and tips to reduce plastic usage. The fake news statements included in the treatment group flyers were drawn from a pool of fake news stories fact checked for accuracy by AltNews.in and BoomFactCheck.com. In order to determine the partisan slant of these fake news stories, I pre-tested each statement on Mechanical Turk with a sample of 600 respondents. For every statement, respondents were asked to determine whether it was favorable to the BJP, not favorable to the BJP, or neutral. News stories that received over 90% consensus from the online sample were picked for the flyer.

The image in Figure 3 is the pro-BJP treatment flyer distributed to respondents.[18] All the fake news statements in this flyer were identified as being favorable to the BJP, and have emanated from government-sympathetic sources (and hence their corrections could be perceived as anti-BJP). Visual cues in the flyer make the partisan identity apparent, for example, the attribution of one piece of fake news to the prime minister, and the use of nationalistic rhetoric. I use other visual cues such as the word 'fake' on the photoshopped images so as to deliver a strong treatment, along with using the color red to underscore corrections.

The Treatment 2 flyer (not pictured) contains fake news messages that are anti-BJP. In order to ensure a balance across flyers in intensity, I made sure to pick similar stories for both flyers (both have a story on Balakot; stories had similar levels of pre-tested salience). The back of both flyers contains tips on how to verify information and spot fake news, and relates back to the stories on the front. The entire intervention was administered in Hindi.

4.4.2 Randomization

In order to control for potential imbalance in the sample, this study uses a randomized block design. The sample was divided into two subgroups based on partisan identity such that the variability

[18] The actual flyer distributed to respondents is in Hindi

within each subgroup was less than the variability between them. Respondents were asked about their partisan identity as a pre-treatment covariate. Specifically, they were asked which national political party in India they feel closest to. Those respondents who choose the BJP were one block, those who choose anything but the BJP were another block. Within each block, respondents were randomly assigned to one of the three experimental groups. This design ensures that each treatment condition has 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 group).

4.5 Hypotheses

Since the primary goal of this study is to test whether media literacy improves ability to identify fake news, the main comparison of interest is between receiving the media literacy training and not receiving it, rather than between the types of partisan stories that were fact-checked. I summarize core hypotheses and predictions below.

Table 2: Main Effects

	Literacy + Pro-BJP fact-check	Literacy + Anti-BJP fact-check
Treatment	A	B
Control	C	C
Difference	$A - C$	$B - C$

Hypothesis 1: *Exposure to the media literacy intervention increases ability to identify fake news.*

There will be a positive effect of the literacy intervention for respondents assigned to any arm of the treatment group relative to placebo control (Groups A + B vs. Group C).

Hypothesis 2a: *Exposure to media literacy and fact-checking pro-BJP fake stories will increase ability to identify fake news.*

Hypothesis 2b: *Exposure to media literacy and fact-checking anti-BJP fake stories will increase ability to identify fake news.*

Additionally, the individual effect of being assigned to each treatment will have a positive effect relative to placebo control (Group A vs. Group C, and Group B vs. Group C such that $A - C > 0$ and $B - C > 0$)

Hypothesis 3: *Ability to identify fake news will be lower among those assigned to receive the literacy and pro-BJP fact-check relative to those assigned to receive the literacy and anti-BJP fact-*

check.

Apart from the effect of receiving any media literacy intervention relative to control, a secondary comparison of interest is between the two treatment arms. Since these treatment groups both received the same media literacy training, they differ only in that the fact-checking for one group involved correcting pro-BJP false stories and the fact-checking for the second group involved correcting anti-BJP fake stories.

Sinha et al (2019) analyse the nature of viral fake news stories in India and demonstrate that the the most salient stories tend to be those favoring the BJP. They also comment on the BJP's stronger hold in the social media realm which contributes to pro-BJP false stories being shared at higher rates. Hence I hypothesize that ability to identify fake news is lower among those assigned to the pro-BJP correction treatment arm compared to the anti-BJP correction treatment arm, relative to control (i.e. $|B - C| > |A - C|$). This prediction is consistent with pre-tests on Mechanical Turk demonstrating that within a large story pool, respondents were more familiar with pro-BJP stories and admitted to having seen them more often on social media relative to anti-BJP stories. The prediction is also consistent with research documenting that information about salient political disputes is processed differently from non-salient information: willingness to engage in effortful resistance to corrections may vary depending on whether the issue is salient and well-known (Jerit and Barabas, 2012).

Hypotheses 1-3 summarize average treatment effects (ATE). Beyond the ATE, I expect the treatment effect to differ conditional on a single factor previously identified in the literature as arguably the most significant predictor of information consumption: partisan identity. The hypothesis for heterogeneous effects with party identity is specified below.

Hypothesis 4: *Effectiveness of the intervention will be higher for politically incongruent information compared to politically congruent information, relative to the control condition.*

Given that the vast literature on directionally motivated reasoning posits that the most common sources of directional motivated reasoning are partisanship and strength of party identity (Nyhan and Reifler 2010), I expect that the treatment effect will be larger for politically incongruent information as compared to to politically congruent information, relative to the control condition. People, especially strong partisans, have difficulty accurately updating their beliefs after finding out that information they previously accepted has been discredited, and hence belief in the most significant fake news stories and misperceptions is often quite stable over time (Nyhan, 2012). In this study, the politically congruent condition manifests when a BJP partisan is provided with anti-BJP fake news that is fact-checked (i.e., the corrected information is now in the favor of the BJP). Similarly, a politically incongruent information condition manifests when a BJP supporter is provided with pro-BJP information that is fact-checked (i.e. the corrected information is now not in the favor of the BJP). Hence this paper predicts that

$$|B - C| > |A - C| \quad \text{among BJP partisans}$$

$$|A - C| > |B - C| \quad \text{among non-BJP partisans}$$

4.6 Dependent Variables

Table 3: 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

In the endline survey, enumerators revisited the same respondents after they had voted in the 2019 elections. The same set of enumerators conducted the baseline survey, intervention, and the endline survey. However, enuemerators were given a random set of household addresses for the endline survey so as to minimize the possibility the same enumerator systematically interviewing the same respondent twice. Further, addresses and contact information were separated immediately from respondent 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 a low attrition rate of 6.2%. Because we were able to match back attrited respondents' contact information to assigned treatment group, the attrition rate was 4% among those assigned to treatment and 9% within control. Importantly, these respondents did not take the endline survey because enumerators did not find them at home; not one respondent refused to answer the endline survey.

The goal of the endline survey was to measure whether respondents were better at identifying fake news as a function of having received the media literacy intervention. To this end, respondents were shown a series of fourteen news stories of which 12 were false and two were true. These stories varied in content, with some focusing on Pakistan-related foreign affairs and some focusing on domestic politics. The stories varied in salience with some that went viral and some that did not. They also varied in visual presentation, with some having accompanying photos and some just text. Lastly, they varied in terms of partisan slant, with half (six) of the false stories having a pro-BJP slant and the other half having an anti-BJP slant[19]. Each respondent saw all the fourteen stories, but the order in which they were shown was randomized to control for order effects.[20] Table 3 summarizes the fourteen stories shown to respondents.

The key dependent variable measure was aimed at measuring the actual efficacy of the intervention. This was measured in two ways immediately following each story:

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

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

After the fourteen political stories, two additional dependent variables were measured: self-perceived efficacy of the treatment, and self-reported media literacy. Self-perceived efficacy was measured by asking respondents "How confident are you that you can spot fake news from real news?" (4-point scale, 1 = very confident, 4 = not confident at all). Media literacy was measured in two ways: respondents were asked how much they trusted news received over WhatsApp (4-point scale); and how frequently they forwarded political messages over WhatsApp (6-point scale). The rationale behind measuring self-reported literacy and efficacy measures is to determine not only whether the intervention was effective at increasing actual ability to spot fake news, but also whether it was successful at generating awareness of the problem, arguably demonstrated by decreased trust in WhatsApp and forwarding of political stories. Research demonstrates that attention to news is lower on mobile phones as compared to other mediums (Dunaway et al, 2018), hence it is important to be able to measure whether the intervention increased attention to news received over WhatsApp thereby decreasing the circulation of false 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.[21] Because respondents were surveyed within 1-3 days of having voted, the presence of an inked finger is a clean and perfect measure of voter turnout.[22]

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

[20] 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 fixed 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 in Appendix Tables 16 and 17.

[21] Sinha, Shishir. (2019). The ink which marks indelibly. *The Hindu BusinessLine*, May 13. https://www.thehindubusinessline.com/opinion/columns/the-ink-which-marks-indelibly/article27119280.ece

[22] Though this may not be true for instances where respondents refuse to show their ink, in this study every

5 Data

In the baseline survey with respondents, enumerators asked questions about several pre-treatment and demographic variables measured in the survey that are likely to affect the treatment effects of this experiment. The literature on misinformation points to several variables having a moderating effect on vulnerability to fake news. Pre-treatment covariates include digital literacy, political knowledge, trust in the news media, and priors about fake news. Demographic variables include age and education.

Following from Zaller (1992), there is reason to believe that political knowledge has an impact on directionally motivated reasoning. Respondents with high levels of political knowledge have strong prior beliefs and those with low levels of political knowledge do not have enough of an interest in politics to become victims of fake news; hence those with an average or middle level of political knowledge are the most vulnerable. In this study political knowledge is measured pre-treatment by a battery of 6 questions of varying difficulty on local and national politics in India. The literature on misinformation has also demonstrated that because individuals have limited time and cognitive resources to comprehend complex topics such as policy or current affairs, they may use the perceived credibility of sources as a heuristic to guide their evaluation of what is true or false. Swire at al. (2018) posit that trustworthiness is a key component of source credibility and suggest that an effective way to reduce misperceptions is to increase the trustworthiness of the source. Because most false news stories in India are without a source, one can think of the medium of information as a proxy for the source of information. To this end, I measure respondent trust in news received over television, in print, and on WhatsApp.

An important moderator of vulnerability to misinformation in the American politics literature is digital literacy. Guess, Nagler and Tucker (2018) demonstrate that digital literacy strongly predicted the propensity to consume fake news during the 2016 US elections; Munger et al. (2018) show that people who are less digitally literate are more likely to fall for fake news and clickbait. However, there is little theoretical clarity for how these findings map on to the Indian context. Because fake news in India is shared frequently on WhatsApp, a certain amount of digital literacy is required in order to broadcast a message, forward it to a list, create a new group etc. Thus for the Indian context, it is possible that higher digital literacy is associated with a lower vulnerability to fake news. But it is also possible that the relationship between perceptions of accuracy of fake news and digital literacy is a U-shaped curve: those with little digital literacy arguably do not use WhatsApp much and those with a lot of digital literacy have the tools to fact-check and tell real from fake, so it is likely those with a middling level of digital literacy who are the most vulnerable to fake news. In this paper, digital literacy is measured through eight five-point (self-reported) ratings of degree of understanding of WhatsApp-related items (adapted for the Indian context from Hargittai, 2005).

Lastly, to measure prior beliefs about fake news, respondents are given a battery of four

respondent willingly showed enumerators their index finger and no one refused.

questions (two true and two false) and for each are asked to determine whether the story is fake or not.

With respect to demographic variables, Lazer (2018) shows that age is an important factor in determining vulnerability to fake news, in that people over the age of 65 are significantly more likely to fall for false and misleading stories. Though age might prove a significant predictor of outcomes in the India context, it is important to note that the sample for this paper skews heavily in towards a younger age cohort. About 83% of the sample was below the age of 35 in this study, pointing to the user profile of those who have their own cellphone with WhatsApp and working internet. Ex ante, it is unclear whether age will have an effect on outcomes in this context.

Religion and caste are important indicators of identity in the Indian context. The National Election Studies (2019) demonstrate that Muslim voters in India are more likely to oppose the BJP, due to the nature of the BJP's recent anti-minority and anti-Muslim politics and rhetoric.[23] Hence it is possible that Muslim voters will more likely believe fake news that is anti-BJP in nature. According to the 2011 Indian census, Muslim voters are also more likely to be poorer as compared to Hindu voters, hence we might see a heterogeneous effect for income wherein lower levels of income translate to a higher resistance for anti-BJP fake news corrections. However education may bring media exposure and awareness, making people more likely to understand the difference between fake and real news. Further, treatment effects may be a function of gender. In a patriarchal society such as India, men are more likely to be educated as compared to women, have more income, own a cellphone (Lokniti-CSDS, 2019). There is also some suggestive evidence has also shown that men in societies with strong patriarchal roots tend to have stronger opinions, and when given new information are less likely to change (Karpowitz et al., 2012). Hence it is possible that there is a reduced effect of fact-checking on the ability to identify fake news for men (as compared to women), and that men are more resistant to pro-BJP fake news corrections. 90% of the sample in this study is male, underscoring both the profile of the average WhatsApp and internet user in a non-metro city in India, as well as the finding that men are more likely to own a cellphone.

There is little direct evidence linking the demographic and pre-treatment variables listed above to the identification and consumption of fake news; specifically, no study in the Indian context provides evidence demonstrating that media effects and information consumption are a function of these variables. Hence though it is possible that treatment effects will differ conditional on these factors, the theoretically ambiguous nature of these variables suggests that the findings from any tests including them will only be speculative. Appendix table 9 presents descriptive statistics for all variables analyzed in this paper.

5.1 Descriptive Analysis

I first begin with descriptive analyses that paint a picture of the extent of belief in fake news, as well as partisan polarization in this belief.

Figure 4 lists the 12 false rumors used as dependent variable measures in this study. This

[23]Lokniti-CSDS 2019 election studies. Data retrieved from https://www.lokniti.org/national-election-studies

figure plots the share of respondents in the sample who believed each story to be true. Three aspects of the figure are striking. First, we notice that general belief in misinformation is low; the sample of respondents in this study appears to be more informed than one would expect given pervasiveness and salience of these rumors. Second, belief in pro-BJP fake news appears to be stronger than belief in anti-BJP fake news. This is evidenced by the fact that top three stories that were believed to be fake were all pro-BJP rumors. This finding supports Jerit and Barabas' (2012) argument that salience increases belief in information, given that Sinha et al. (2019) document how the right-wing fake news machinery in India is stronger and more pervasive, leading to higher salience. Third, belief in Pakistan and foreign terrorism-related fake news (Pulwama Photo, Terror Attacks, BJP Soldier) is higher as compared to domestic issues, and belief in election tampering rumors is low overall (EVM Rigged, Poll Booth, Duplicate Vote). Lastly, the highest levels of belief (47.54%)were reserved for the "Gomutra" rumor, the story that cow urine can cure cancer.

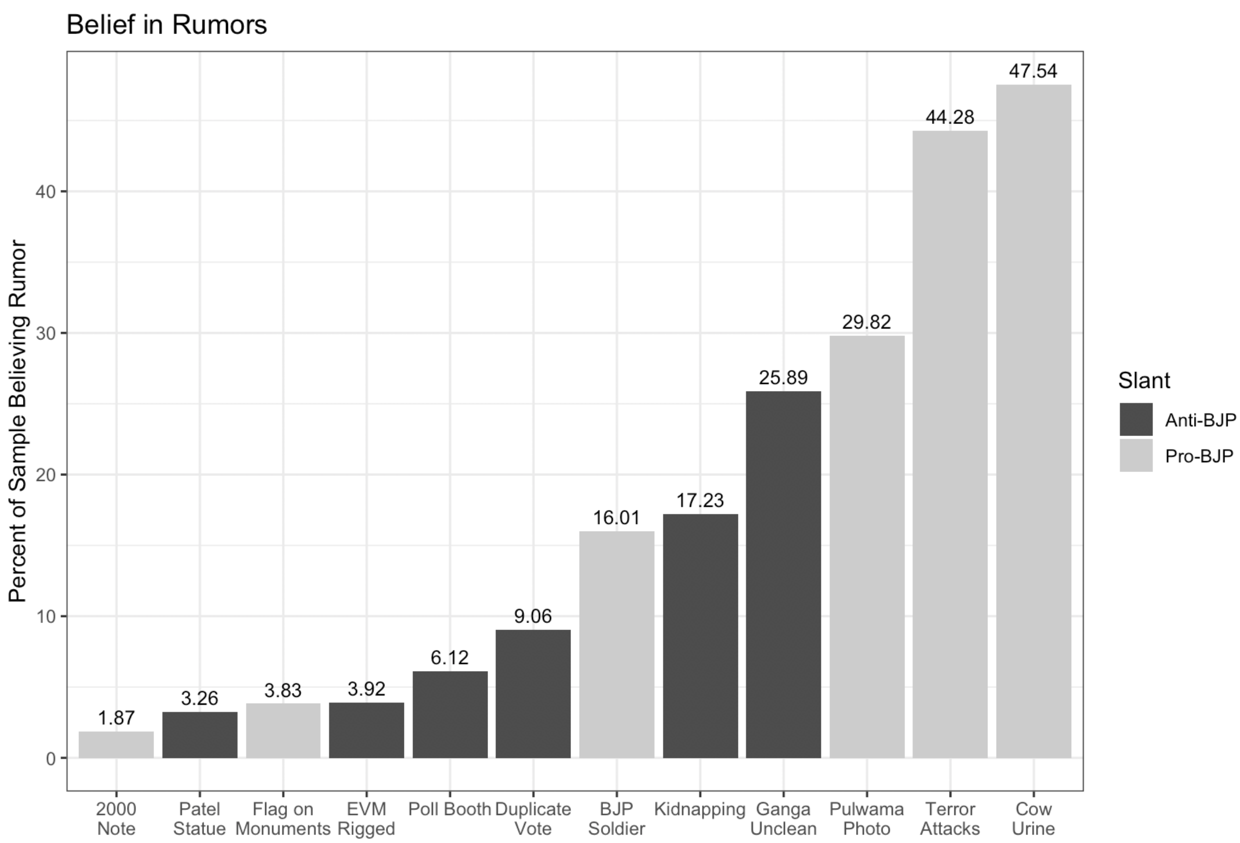


Figure 4: Percent of Sample Who Believe Rumors

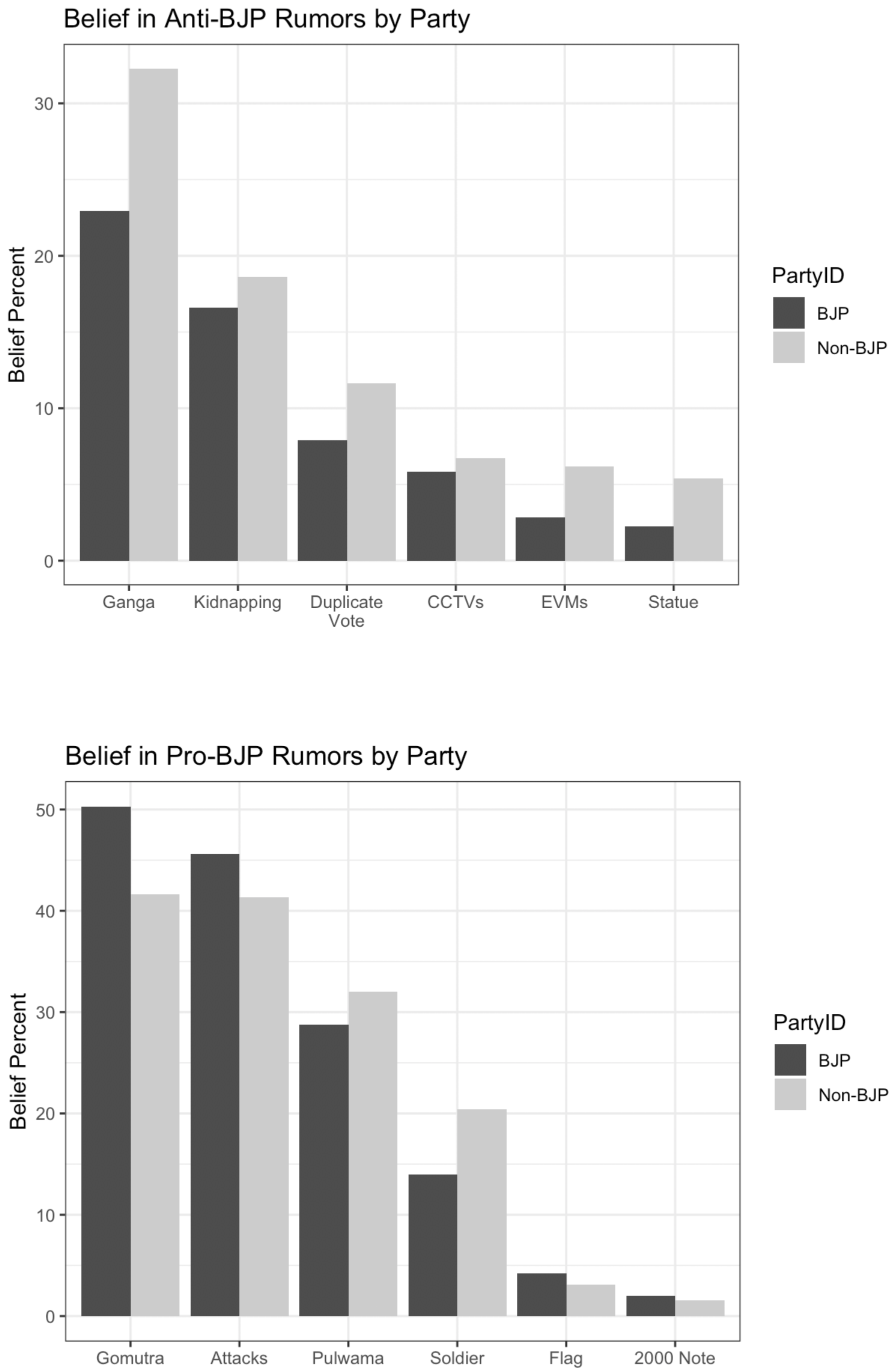


Figure 5: Belief in Rumors by Respondents' Party ID

In Figure 5 I plot respondent belief in rumors by their self-reported partisan identity. For 10 out of the 12 partisan rumors, we see a correspondence between respondent party identity and pre-tested political slant of the rumor. An important observation here is that misinformation does not seem as politically polarizing in India as it is in the United States. 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).

6 Estimation and Results

The last part of the intervention involved enumerators administering a comprehension check aimed at measuring whether the treatment was effective in the short-term i.e. in immediately reducing the perceived accuracy of the fake stories that were debunked in the flyer. Respondents were provided with two false stories that were debunked by enumerators in the same house visit: (1) the new 2000 rupee note issued by the BJP has a tracking chip in it so as to reduce the circulation of black money, (2) the new Patel statue built by the government with taxpayer money has already developed cracks underscoring infrastructural shortcomings of the government. For each story respondents were asked to identify whether it was fake or not. For story (1), only 5.83% of the sample said that the story was true; for story (2), only 3.40% of the sample said that the story was true. This demonstrates that in the short run, respondents were able to successfully identify stories as fake after they had been debunked. Next respondents were asked how they knew the story to be true or false. This was an open ended question aimed at understanding whether it was prior knowledge about the story or whether the current intervention helped respondents in its identification as fake or not fake. About 57% of the sample said that they were able to identify the story as being true or false because of the debunking in the intervention (their response translated to "you just told us about this story") relative to the remaining 43% of the sample who stated reasons such as seeing the story on the news, reading it in the papers, or hearing about it from a family member. Taken together these results demonstrate that respondents understood the treatment.

I estimate effects of the treatment on outcomes in a between-subjects design: treatment assignment (all treatment group respondents versus placebo control), and type of treatment (each flyer versus placebo control). All estimates are ordinary least square (OLS) regressions in which the treatment variable is either a pooled indicator for assignment to any treatment group relative to control, or an indicator for assignment to one treatment group. For all analyses I specify empirical models relying on random treatment assignment to control for potential confounders.

I estimate the following equations to test the main effect of the media literacy intervention on ability to identify fake news:

$$FakeNewsId_i = \alpha + \beta_1 Treatment_i + \epsilon_i \tag{1}$$

$$FakeNewsId_i = \alpha + \beta_1 LiteracyPro\text{-}BJP_i + \beta_2 LiteracyAnti\text{-}BJP_i + \epsilon_i \tag{2}$$

In the equations, i represents the respondent, the *Treatment* variable in Equation 1 represents pooled assignment to media literacy intervention (relative to control). In Equation 2, the dependent variable is regressed on separate indicators for having received the media literacy and pro-BJP flyer or media literacy and anti-BJP flyer, with the control condition as the omitted category. The dependent variable *FakeNewsId*, representing the respondent's ability to identify stories as fake, measures the raw count of the number of stories correctly identified as fake. This variable is constructed such that it takes on the value of 1 if a respondent correctly identifies a given story to be fake. The maximum value for the variable is 12 and the minimum is 0. *FakeNewsId* has been coded such that a positive estimated β_1 indicates an increase in the ability to identify fake news.

A common finding in the literature on factual beliefs and misinformation is that partisan attachment is an important predictor of beliefs. To determine whether partisan identity moderates treatment effects, I next interact treatment indicators with partisan identity. This is done in two ways as indicated by the equations below:

$$FakeNewsId_i = \alpha + \beta_1 Treatment_i + \beta_2 Treatment_i * PartyID_i + \beta_3 PartyID_i + \epsilon_i \tag{3}$$

$$FakeNewsId_i = \alpha + \beta_1 CongruentTreatment_i + \beta_2 IncongruentTreament_i + \epsilon_i \tag{4}$$

In Equation 3, i represents a respondent and the *Treatment* variable again represents assignment to any treatment arm relative to control. *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 the party identity variable as dichotomous was based on the nature of misinformation in India where false stories are perceived as either favoring or not favoring the BJP. Further, the sample for this paper was overwhelmingly supportive of the BJP at the national level, with only 45% of respondents identifying as supporters of parties other than the BJP. Hence I code party identity as binary support for the BJP relative to all other parties. In this equation, the interaction term β_2 is the coefficient of interest. A positive coefficient estimate for β_2 indicates an increase in the ability to identify fake stories among BJP partisans due to the treatment.

In Equation 4 I determine whether partisan fact-checking in the treatment affects outcomes. To do this the equation estimates whether BJP supporters are better at identifying fake news when they receive a partisan-congruent treatment. In this experiment, a partisan-congruent treatment manifests when corrections to fake news are pro-party. Hence for a BJP partisan, reading anti-BJP stories which are then corrected manifests in a congruent condition, because the corrections are pro-BJP. Similarly, when a BJP partisan reads pro-BJP fake stories that are fact-checked, the corrections become anti-BJP resulting in an incongruent condition. To estimate the effect of partisan fact-checking on BJP respondents, I construct the variable *CongruentTreatment* that takes the value of 1 if the respondent said they were a BJP supporter and if they received pro-BJP corrections. The variable *IncongruentTreatment* takes on the value of 1 if the respondent is a BJP supporter and received anti-BJP corrections.

First, I analyse data on the main effect of the intervention on ability to identify fake news.

While research predicts that in-person and field interventions on media effects are likely to have stronger effects, my findings from misinformation-prone India are less encouraging. Even with an in-person intervention, where enumerators spend close to one hour with each respondents to debunk and discuss misinformation and where respondents understood the intervention, I do not see significant increases in the ability to identify fake news as a function of teaching respondents media literacy tools.

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

Table 4: Effect of Treatment on Ability to Identify Fake News

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

Table 4 clearly indicates that the literacy intervention did not increase fake news identification ability in any meaningful way. Analysis by splitting the treatment into its component parts (each compared to placebo control) yields similar results.

In summary, I find no evidence that an hour-long media literacy intervention increased ability to identify fake news among respondents connected to the internet in Bihar, India, and I find no significant effects associated with being assigned to receive a partisan fact-checking along with literacy training. Vulnerability to fake news remains entrenched and largely unaffected by any aspect of the in-person intervention. As demonstrated by the low r^2 value in these models, it appears that the literacy intervention was not successful at explaining the dependent variable as there is virtually no evidence that such an intervention has desirable effects on ability to identify false stories after a period of time. The ability to update one's priors in response to factual information is privately and socially valuable, and hence the fact that a strong, in-person treatment does not change opinions demonstrates the stickiness of fake news in India. Priors about fake news in this study context appear resistant to change but, as I demonstrate below, this does not preclude moderating effects of partisan identity.

Next, to test whether ability to identify fake news is different for respondents assigned to one treatment arm as compared to the other, I run tests to assess the joint significance of predictors in the main ATE model. To do this, I estimate a model where I assume that the Pro-BJP Fact-check coefficient is the same as the Anti-BJP Fact-Check coefficient. The equation below specifies this model:

$$FakeNewsId_i = \alpha + \beta_1(LiteracyPro\text{-}BJP + LiteracyAnti\text{-}BJP)_i + \epsilon_i \qquad (5)$$

I then use an Anova test to determine the difference between the models in equation 2 and 5. Results are in Table 5. The p-value of 0.66 highlights that the two models are not significantly different from each other, demonstrating that there was no statistical difference in ability to identify fake news among those assigned to receive the anti-BJP fact-checking treatment relative to those assigned to receive the pro-BJP fact-checking treatment.

Table 5: Anova Output for Hypothesis 3

Statistic	Mean	St. Dev.
Res.Df	1221.5	0.707
RSS	3005.01	0.394
Df	−1.000	
Sum of Sq	−0.558	
F	0.227	
Pr(>F)	0.634	

I now turn to the analysis of heterogeneous effects of partisan identity. A common finding in the literature on factual beliefs and misinformation is that partisan attachment is an important predictor of beliefs and research has routinely demonstrated that receiving an intervention could affect the way certain partisans update their priors. If these findings were to carry over to the Indian context, what kind of fake stories are BJP partisans more or less likely to believe after having received a strong treatment? Table 6 presents results. In column 1 I estimate the effect of receiving the treatment for BJP supporters on the ability to identify pro-BJP fake stories, column 2 does the same with anti-BJP fake stories. The treatment variable for both models pools across receiving any treatment relative to control.

Table 6: Effect of Treatment x Party on Ability to Identify Fake News

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

Results are striking: while there was no average treatment affect, the interaction effect for the treatment on BJP partisans produces a negative effect on the ability to identify fake news – i.e.the treatment is less effective on BJP partisans – but only for pro-BJP stories. This effect does not produce for anti-BJP stories for which BJP partisans seem better at identifying fake news.

Does partisan fact-checking in the treatment affect outcomes? In order to answer this question, I determine whether BJP supporters are better at identifying fake news when they receive a partisan-congruent treatment. In this experiment, a partisan-congruent treatment manifests when corrections to fake news are pro-party. Hence for a BJP partisan, reading anti-BJP stories which are then corrected manifests in a congruent condition, because the corrections are pro-BJP. Similarly, when a BJP partisan reads pro-BJP fake stories that are fact-checked, the corrections become anti-BJP resulting in an incongruent condition. To estimate the effect of partisan fact-checking on

BJP respondents, I construct a variable that takes the value of 1 if the respondent said they were a BJP supporter and if they received pro-BJP corrections (congruent condition). The variable thus takes on the value of 0 if the respondent is a BJP supporter and received anti-BJP corrections (incongruent condition).

Results are presented in Table 7. Consistent with the party effects reported above, BJP partisans are more likely to identify stories as fake when those stories are anti-BJP in nature, but they do so only when receiving the congruent treatment. The magnitude of this effect is smaller than for the partisan effect reported in Table 5 (p=0.73). The same effect for anti-BJP fake stories does not produce when BJP partisans receive the incongruent treatment, demonstrating that although the treatment did not correct all the stories in the DV, receiving a party-consistent treatment primes partisanship in BJP supporters, producing a higher identification for anti-BJP false stories and no effect for pro-BJP false stories.

Table 7: Effect of Party Congruence in Treatment on Ability to Identify Fake News

	Dependent variable: Number of Stories Identified as Fake	
	Pro-BJP Stories	Anti-BJP Stories
	(1)	(2)
Congruent Treatment	-0.137^{*}	0.120^{*}
	(0.079)	(0.066)
Incongruent Treatment	-0.085	0.095
	(0.079)	(0.066)
Constant	4.617^{***}	5.296^{***}
	(0.043)	(0.036)
Observations	1,224	1,224
R^2	0.003	0.003
Adjusted R^2	0.001	0.002
Residual Std. Error (df = 1221)	1.109	0.924
F Statistic (df = 2; 1221)	1.663	2.094
Note:		*p<0.1; **p<0.05; ***p<0.01

While it seems rational for strong partisans to deliberately discredit the out-party whilst simultaneously accepting misinformation that paints their party in a good light, it is worth investigating why this effect might manifest only for treatment group BJP respondents. One explanation for these results is that in asking respondents questions about particular stories, those stories are made salient in their minds, likely to a greater extent for stories that were salient in the news to begin with (pro-BJP fake news) and for stronger partisans (BJP supporters). Thus the negative coefficient on Treatment x BJP could be an exaggerated response due to partisans' stronger tendency to want to "cheerlead" for their party, as opposed to control group BJP respondents (Bullock, Gerber, Hill, and Huber 2015; Gerber and Huber 2009; Prior, Sood, and Khanna 2015). I explore this result further in the Discussion.

Independent of the literacy intervention, it is descriptively interesting for the understudied context of India to understand who is more likely to consume fake news and more likely to be able to identify news as false. I consider the main effect of several demographic and pre-treatment variables on ability to identify fake news. The results are presented in Table 8. For all 12 false stories taken together, BJP partisans are significantly better at identifying fake news as compared to their non-BJP partisan counterparts. Further, as expected, accurate prior beliefs about fake stories are more likely to aid in identifying fake news. Surprisingly, higher levels of digital literacy were negatively associated with fake news identification. In this study, digital literacy was measured by a battery of 6 items asking respondents about their levels of familiarity with WhatsApp-related terms. Results demonstrate that those with higher levels of digital literacy are less likely to be able to identify fake stories, underscoring that greater knowledge of WhatsApp likely leads to more vulnerability to fake news. In a similar vein, those who report using WhatsApp more often are also less likely to be able to identify fake news. Interestingly, higher levels of trust in WhatsApp do not correlate with identification of fake news stories, suggesting that time spent and familiarity with the medium itself make people more vulnerable to misinformation.

With respect to demographic variables, the correlation most statistically significant is age. While scholarship on misinformation in the United States demonstrates that older respondents are worse at identifying fake news, in this data increasing age is associated with a higher capacity to identify fake news. Further, the association of education with correct beliefs should be highlighted. Flynn, Nyhan, and Reifler (2017) argue that education could have opposing effects on political misperceptions. On the one hand, education should increase people's ability to discern fact from fiction. On the other hand, in the presence of motivated reasoning, education gives people better tools to counter-argue against incongruent information. In this study, education has a positive effect in ability to identify news as fake.

6.1 Discussion

The most striking findings to emerge from this study demonstrate that media literacy training interventions do not affect vulnerability to misinformation in this context, but the treatment backfires for a subset of partisans who are less likely to identify false stories that portray their party in a positive light.

Taken together, these results have two main implications. First, though partisan divides

Table 8: Main Effect of Demographic and Pre-Treatment Variables

	Dependent variable: Number of Stories Identified as Fake
	Pooled DV : All Stories
Literacy Intervention	−0.060
	(0.095)
BJP Supporter	0.234**
	(0.113)
Accurate Priors	0.480**
(Higher = more accurate)	(0.231)
Digital Literacy	−1.168***
(Higher = more literate)	(0.252)
Political Knowledge	−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	10.185***
	(0.396)
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

are expected for political attitudes and tastes,[24] this paper shows that for certain partisans, these divides may extend to factual beliefs. The finding from this paper that certain partisans are significantly less likely to identify stories as false particularly when those stories depict their party in a positive manner demonstrates that for BJP supporters, partisanship can be a perceptual screen (Campbell, 1960). Hence divergences in factual beliefs are a natural consequence of the effect of partisanship on information processing (Gerber, Huber and Washington, 2010). Further, given the expectation of increased BJP campaign salience, this finding supports research demonstrating that citizen attachments to political parties are hightened during elections (Michelitch and Utych, 2018). Second, since some of the strongest defenses of democracy rely on theories of retrospective voting, where citizens can administer justice by rewarding or punishing incumbents for for things that happened during their terms (Fiorina 1981), partisan differences to factual stories call into question voters' abilities to vote retrospectively. Partisan bias in factual news may simultaneously provide politicians with weaker incentives to perform better in office and stronger incentives to spread misinformation that is beneficial to their party.

Research also contends, however, that partisan differences in survey responses may not solely indicate differences in true beliefs but may reflect expressive survey responses to questions that portray one's party in a positive light. Just as people enjoy cheering for their favourite team or player even when they know the limitations of the team, people gain utility from cheerleading for their favourite party, but this does not necessarily mean they do not understand the true merits and demerits of their preferred candidates (Bullock et al, 2013; Green, Palmquist and Schickler, 2002). Hence such research suggests that a viable solution to reducing partisan gaps in survey response might be to design studies with incentives to answer correctly.

However such solutions may point to a deeper problem that presents itself in the Indian case: generation of accurate survey responses may not translate into a decrease in expressing only true factual beliefs in real life, as mediums where misinformation is spread in India such as WhatsApp do not impose costs on the sender for sharing incorrect stories. Thus solutions that monetarity incentivize correct responses in surveys lack external validity when applied to the Indian setting. A logical extension of this solution is to actually impose costs to sharing false messages *within* the medium through which they are shared. This would involve developing an incentive structure such that costs are imposed on the user for sharing messages that are demonstrated to be false, through either group chat platforms completely disallowing indisputably factually false messages from being forwarded in groups, or by limiting the number of times a user can forward a *disputed* message thereby reducing the number of people it is forwarded to. Such solutions may have the double advantage of both reducing vulnerability to fake news on average, as well as disincentivizing expressive partisan behavior for strong partisans.

Taxing social media is an idea that certain emerging economies have already toyed with. In 2018, the Ugandan government introduced an excise duty bill aimed at levying a small daily tax on

[24] Survey research from India has demonstrated a large partisan gap between BJP and other partisans in attitudes towards the favorability of the way in which domestic and international politics issues have been handled by Prime Minister Narendra Modi (Stokes et al, 2017)

users of WhatsApp, among other social media applications.[25] Although government leaders believed the move would help in reducing false news and gossip, in reality its goals were to impose a blanket everyday tax on all users for accessing the application alone. In India, WhatsApp rolled out two measures aimed at curtailing falsehoods from becoming viral: adding a "forwarded" tag to messages that were not composed by the sender, and limiting the number of times one could forward a single message to five. Although these measures are intended to reduce the spread of misinformation, they do not particularly target false stories. Despite the forwarding limit, messages can still be forwarded to groups with each group including up to 256 people, meaning that a forwarded message could be shared with nearly 1,300 people instantly, in spite of the five time limit. The presence of several rumors and false stories that went viral during and before the 2019 election in India, such as the ones tested in this paper, demonstrates that such measures have not been successful at curbing misinformation.

Though studies that test cost-imposing interventions to restrict the spread of falsehoods have not thus far been explored, research has demonstrated the likely benefits of imposing a transaction tax on financial trading in order to "throw sand into the gears" of well functioning financial markets (Tobin, 1982; Summers and Summers, 1989). Akin to this is the spread of WhatsApp usage in India, where prioritizing cheap data rates and access to internet in remote areas has increased development, contributing to the immense popularity of the application. Paradoxically, these same factors have lent themselves to an assault on factually accurate information, creating an environment where the incentives for misinformation to be created and flourish outweigh the costs to curb it.

My results demonstrate that misinformed beliefs are so deep-seated in the Indian electoral context that bottom-up, strong, in-person fixes fall short of fixing the problem. Hence if interventions targeted at user-training do not work, perhaps the solution lies in top-down fixes to the system itself. Though implementing legal solutions that impose costs is a long-term and time consuming process, future research will benefit from analysing whether experimental studies that imposing costs within chat environments are likely to produce desired behaviors.

7 Conclusion

The aftermath of the 2016 American presidential election saw a surge of research and scholarship on fake news and its effects. But misinformation has been a growing concern across the world and particularly in developing countries, where false claims can distort public opinion, undermine faith in democracy, and endanger public health. In this paper, I present new evidence on belief in popular fake news stories in India in the context of the 2019 general elections. Given the seriousness of the problem of fake news and the lack of means to measure its pervasiveness, I run a field intervention designed to increase media literacy in treatment group respondents in a challenging context, where both misinformation is hightened during elections and where false stories have often been linked to violence and murder.

[25]BBC News (2018). Uganda imposes WhatsApp and Facebook tax 'to stop gossip. BBC, May 31.https://www.bbc.com/news/world-africa-44315675

I find that a grassroots-level intervention in which respondents spend an hour with a survey enumerator and learn about concrete tools to fact-check misinformation has little effect on respondent ability to identify fake news. These results are striking in that they demonstrate that while subtle treatments such as "disputed" tags on messages produce effects in the American context, the hour-long in-person fact-checking and media literacy in my study falls short in its goal of changing priors on fake beliefs. Overall I find no average treatment effect across individual fake stories or a pooled DV of stories. But, an interesting partisan effect manifests: the partisanship and polarization of BJP supporters appears stickier than that of their out-partisans. BJP partisans better overall at identifying fake news, but conditional on receiving the treatment their identification ability differs. When it comes to identifying pro-attitudinal stories as fake, BJP supporters who receive the treatment respond in a way that is the opposite of what is hypothesized – treatment makes them less likely to identify stories as fake. Important to note is that this effect only produces for BJP supporters when it comes to fake news that is pro-BJP in nature, underscoring the possibility of a more consolidated pro-BJP party identity that moves respondents to answer expressively or cheerlead for their party.

To the extent that consuming misinformation imposes costs and given that a strong, in-person treatment does not bring desired results, what can be done? One possibility is to design a study with incentives for correct responses, so as to control for partisan gaps. However, as most fake information is circulated over WhatsApp with no cost on the sender for sharing it with several hundred of his or her contacts, this study points to an even deeper problem: generation of an accurate survey response may not translate into a shift in true opinion or behavior in the current context.

These findings suggest that fake beliefs are entrenched to the extent that educative and training interventions do not move these beliefs. Hence future research should look at longer training sessions that are not one-shot interventions so as to educate respondents in the long run, as well as studies that test the effect of imposing a cost on sharing a false message. Such studies should not overlook the role of partisan identity in India given that these results demonstrate a more consolidated BJP party identity relative to conventional wisdom.

8 References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211-236.

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4), 619-638.

Bolsen, T., & Druckman, J. N. (2015). Counteracting the politicization of science. *Journal of Communication*, 65(5), 745-769.

Brown, K., Campbell, S. W., Ling, R. (2011). Mobile phones bridging the digital divide for teens in the US? *Future Internet*, 3(2), 144–158

Bullock, J. (2007). Experiments on partisanship and public opinion: Party cues, false beliefs, and Bayesian updating. Ph.D. dissertation, Stanford University.

Chandra, K. (2004, July). Elections as auctions. In Seminar (Vol. 539).

Chhibber, P. K., & Verma, R. (2018). *Ideology and identity: The changing party systems of India.* Oxford University Press.

Clayton, K., Blair, S. Forstner, S., Glance, J., et al. (2019). Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Banners in Reducing Belief in False Stories on Social Media, *Political Behavior*, 32(2), 207-218

Constine, J. (2017). Facebook puts link to 10 tips for spotting false news atop feed. Tech Crunch, April 6, 2017. Retrieved July 18

Devlin, K. and Johnson, C. (2019). elections nearing amid frustration with politics, concerns about misinformation. Pew Research Center.

Druckman, James N. (2012). The politics of motivation.*Critical Review*,24 (2): 199-216.

Druckman, James N., Erik Peterson, and Rune Slothuus. (2013). How Elite Partisan Polarization Affects Public Opinion Formation. *American Political Science Review*, 107 (1): 57-79.

Dunaway, J., Searles, K., Sui, M., and Paul, N. (2018). News Attention in a Mobile Era. *Journal of Computer-Mediated Communication*, 23(2), 107-124.

Donner, J., Walton, M. (2013). Your phone has internet-why are you at a library PC? Re-imagining public access in the mobile internet era. In P. Kotze, et al. (Eds.), Proceedings of INTERACT 2013: 14th IFIP TC 13 International Conference (pp. 347–364). Berlin: Springer.

Ecker, U. K. H., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied*, 20(4), 323-335.

Flynn, D. J. (2016). The scope and correlates of political misperceptions in the mass public. In *Annual Meeting of the American Political Science Association, Philadelphia.*

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38(S1), 127-150.

Freed, G. L., Clark, S. J., Butchart, A. T., Singer, D. C., & Davis, M. M. (2010). Parental vaccine safety concerns in 2009. Pediatrics, 125(4), 654–659.

Fridkin, K., Kenney, P. J., & Wintersieck, A. (2015). Liar, liar, pants on fire: How fact-checking influences citizens' reactions to negative advertising. *Political Communication*, 32(1), 127-151.

Gitau, S., Marsden, G., Donner, J. (2010). After access: Challenges facing mobile-only internet users in the developing world. In Proceedings of the 28th International Conference on Human Factors in Computing Systems (pp. 2603–2606). New York: ACM.

Gentzkow, Matthew A., and Jesse M. Shapiro. (2004). Media, Education and Anti- Americanism in the Muslim World. *Journal of Economic Perspectives*, 18 (3): 117-133.

Gentzkow, M., Shapiro, J. M., & Stone, D. F. (2015). Media bias in the marketplace: Theory. In Handbook of media economics (Vol. 1, pp. 623-645). North-Holland.

Gottfried, J., & Shearer, E. (2017). *News use across social media platforms 2016.* Pew Research Center, May 26, 2016

Guess, A., Nagler, J., & Tucker, J. A. (2018). Who's Clogging Your Facebook Feed? Ideology and Age as Predictors of Fake News Dissemination During the 2016 US Campaign. Unpublished manuscript.

Hargittai, E. (2005). Survey measures of web-oriented digital literacy. *Social science computer review*, 23(3), 371-379.

Jerit, Jennifer, and Jason Barabas. (2012). Partisan perceptual bias and the information environment. *Journal of Politics*, 74 (3): 672?684.

Jerit, J., Barabas, J., & Clifford, S. (2013). Comparing contemporaneous laboratory and field experiments on media effects. *Public Opinion Quarterly*, 77(1), 256-282.

Kaka et al. (2019). *Digital India: Technology to transform a connected nation.* McKinsey Global Institute.

Karpowitz, C.F., Mendelberg, T. & Shaker, L. (2012). Gender inequality in deliberative participation. *American Political Science Review*, 106: 533?547.

Kitschelt, H., & Wilkinson, S. I. (Eds.). (2007). *Patrons, clients and policies: Patterns of democratic accountability and political competition.* Cambridge University Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, 108(3), 480.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. *Science*, 359(6380), 1094-1096.

Lau, Richard R., and David P. Redlawsk. (2001). Advantages and disadvantages of cognitive heuristics in political decision making. *American Journal of Political Science*,45 (4): 951-971.

Michelitch, K., & Utych, S. (2018). Electoral cycle fluctuations in partisanship: Global evidence from eighty-six countries. *The Journal of Politics*, 80(2), 412-427.

Mosseri, A. (2017). A new educational tool against misinformation. Facebook, April 6, 2017. Retrieved May 23, 2017

Munger, K., Luca, M., Nagler, J., & Tucker, J. (2018). The Effect of Clickbait. Working Paper.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303-330.

Nyhan, Brendan, and Jason Reifler. (2012). Misinformation and Fact-checking: Research Findings from Social Science. New America Foundation Media Policy Initiative Research Paper.

Nyhan, Brendan, and Jason Reifler. (2013). Which Corrections Work? Research results and practice recommendations. New America Foundation Media Policy Initiative Research Paper

Nyhan, B., Porter, E., Reifler, J. & Wood, T. (2017). Taking corrections literally but not seriously? The effects of information on factual beliefs and candidate favorability. Unpublished manuscript. Retrieved July 8, 2018

Owen, L.H. (2018). Is your fake news about immigrants or politicians? It all depends on where you live. Nieman Journalism Lab, May 25, 2018. Retrieved July 18, 2018

Pennycook, G., & Rand D.G. (2018). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition.* Retrieved July 8, 2018.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General.*

Schaedel, S. (2017). Black lives matter blocked hurricane relief? Factcheck.org, September 1, 2017. Retrieved September 26, 2017

Silver, L. and Smith, A. (2019). *I*n some countries, many use the internet without realizing it. Pew Research Center.

Silverman, C. (2016). This analysis shows how viral fake election news stories outperformed real news on facebook. *Buzzfeed*, November 16, 2016. Retrieved May 22, 2017.

Singh, S. (2019). How To Win An Indian Election: What Political Parties Don't Want You to Know. Mumbai: Penguin

Sinha, P., Sheikh, S., & Sidharth, A. (2019). *India Misinformed.* Noida: Harper Collins.

Silverman, C., & Jeremy S.-V. (2016). Most Americans who see fake news believe it, new survey says. December 6, 2016. Retrieved May 23, 2017

Smith, J., Jackson, G., & Raj, S. (2017). Designing against misinformation. Medium, December 20, 2017. Retrieved July 8, 2018

Stroud, N. (2008). Media use and political predispositions: Revisiting the concept of selective exposure. *Political Behavior*, 30 (3): 341-366.

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. (2017). Processing political misinformation: comprehending the Trump phenomenon. *Royal Society Open Science,*4(3), 160802.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755-769.

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, 33(3), 460-480.

Weeks, B. E., Gil de Zuniga, H. (2019). What's Next? Six Observations for the Future of Political Misinformation Research. *American Behavioral Scientist*

Zaller, J. R. (1992). *The nature and origins of mass opinion.* Cambridge university press.

Supplementary Appendix

Table 9: 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	0.000	0.750	1.000
Digital Literacy	1,224	0.758	0.194	0.083	0.833	1.000
Political Knowledge	1,224	5.000	1.135	0	5	6
WhatsApp Use Frequency	1,224	1.932	0.952	1	2	7
Trust in WhatsApp	1,224	2.729	0.821	1	3	4
Education	1,224	9.388	2.652	1	9	13
Age	1,224	26.646	9.182	18	24	85
Male	1,224	0.911	0.285	0	1	1
Hindu	1,224	0.837	0.369	0	1	1

Table 10: Effect of Literacy Intervention on Pro-BJP Rumor Identification

	Dependent variable: Pro-BJP Rumors					
	Cow Urine	Attacks	Pulwama	Soldier	Flag	2000 Note
	(1)	(2)	(3)	(4)	(5)	(6)
Literacy Intervention	0.022	0.001	−0.030	0.007	−0.002	−0.002
	(0.030)	(0.030)	(0.028)	(0.022)	(0.012)	(0.008)
Constant	0.510***	0.557***	0.722***	0.835***	0.963***	0.983***
	(0.025)	(0.025)	(0.023)	(0.018)	(0.010)	(0.007)
Observations	1,224	1,224	1,224	1,224	1,224	1,224
R^2	0.0004	0.00000	0.001	0.0001	0.00003	0.0001
Adjusted R^2	−0.0004	−0.001	0.0001	−0.001	−0.001	−0.001
Res. Std. Error (df = 1222)	0.500	0.497	0.458	0.367	0.192	0.136
F Statistic (df = 1; 1222)	0.523	0.001	1.146	0.108	0.035	0.079

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11: Effect of Literacy Intervention on Anti-BJP Rumor Identification

	Dependent variable: Anti-BJP Rumors					
	Ganga	Kidnapping	Double Vote	CCTV	EVM	Statue
	(1)	(2)	(3)	(4)	(5)	(6)
Literacy Intervention	−0.008	−0.015	0.008	0.012	0.008	−0.001
	(0.027)	(0.023)	(0.017)	(0.015)	(0.012)	(0.011)
Constant	0.746***	0.837***	0.904***	0.931***	0.956***	0.968***
	(0.022)	(0.019)	(0.014)	(0.012)	(0.010)	(0.009)
Observations	1,224	1,224	1,224	1,224	1,224	1,224
R^2	0.0001	0.0003	0.0002	0.001	0.0003	0.00001
Adjusted R^2	−0.001	−0.0005	−0.001	−0.0003	−0.0005	−0.001
Residual Std. Error (df = 1222)	0.438	0.378	0.287	0.240	0.194	0.178
F Statistic (df = 1; 1222)	0.089	0.410	0.212	0.624	0.422	0.008
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 12: Effect of Literacy Intervention x Party ID on Pro-BJP Rumor Identification

	Dependent variable: Pro-BJP Rumors					
	Cow Urine	Attacks	Pulwama	Soldier	Flag	2000 Note
	(1)	(2)	(3)	(4)	(5)	(6)
Literacy Intervention	0.126**	0.049	0.050	0.052	0.011	−0.012
	(0.054)	(0.053)	(0.049)	(0.039)	(0.021)	(0.015)
BJP Supporter	0.014	0.004	0.111**	0.108***	0.002	−0.014
	(0.053)	(0.053)	(0.049)	(0.039)	(0.020)	(0.014)
Intervention x	−0.153**	−0.071	−0.118**	−0.066	−0.020	0.014
BJP Supporter	(0.065)	(0.065)	(0.060)	(0.048)	(0.025)	(0.018)
Constant	0.500***	0.554***	0.646***	0.762***	0.962***	0.992***
	(0.044)	(0.044)	(0.040)	(0.032)	(0.017)	(0.012)
Observations	1,224	1,224	1,224	1,224	1,224	1,224
R^2	0.011	0.003	0.005	0.008	0.001	0.001
Adjusted R^2	0.009	0.0001	0.003	0.006	−0.001	−0.002
Residual Std. Error (df = 1220)	0.497	0.497	0.457	0.366	0.192	0.136
F Statistic (df = 3; 1220)	4.727***	1.057	2.136*	3.394**	0.492	0.341
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 13: Effect of Literacy Intervention x Party ID on Anti-BJP Rumor Identification

	Dependent variable: Anti-BJP Rumors					
	Ganga	Kidnapping	Double Vote	CCTV	EVM	Statue
	(1)	(2)	(3)	(4)	(5)	(6)
Literacy Intervention	0.046	0.021	0.010	−0.009	0.022	−0.001
	(0.047)	(0.041)	(0.031)	(0.026)	(0.021)	(0.019)
BJP Supporter	0.147***	0.055	0.040	−0.011	0.048**	0.032*
	(0.046)	(0.040)	(0.031)	(0.026)	(0.021)	(0.019)
Intervention x BJP Supporter	−0.081	−0.053	−0.004	0.029	−0.022	−0.001
	(0.057)	(0.049)	(0.037)	(0.031)	(0.025)	(0.023)
Constant	0.646***	0.800***	0.877***	0.938***	0.923***	0.946***
	(0.038)	(0.033)	(0.025)	(0.021)	(0.017)	(0.016)
Observations	1,224	1,224	1,224	1,224	1,224	1,224
R^2	0.012	0.002	0.004	0.002	0.007	0.007
Adjusted R^2	0.009	−0.001	0.001	−0.001	0.005	0.004
Residual Std. Error (df = 1220)	0.436	0.378	0.287	0.240	0.194	0.177
F Statistic (df = 3; 1220)	4.765***	0.765	1.569	0.612	3.000**	2.794**
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 14: 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 15: 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 BJP Supporter	0.039	0.002	0.035	−0.075
	(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

Table 16: Effect of Treatment with Enumerator Fixed Effects

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

Table 17: Effect of Treatment x Party with Enumerator Fixed Effects

	Dependent variable: Number of Stories Identified as Fake	
	Pro-BJP Stories	Anti-BJP Stories
	(1)	(2)
Literacy Intervention	0.254**	0.077
	(0.103)	(0.093)
BJP Supporter	0.265***	0.327***
	(0.102)	(0.092)
Literacy Intervention x 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