# "I Don't Think That's True, Bro!" An Experiment on Fact-checking *WhatsApp* Rumors in India.

Sumitra Badrinathan
University of Pennsylvania

Simon Chauchard
Leiden University

D.J. Flynn
IE University*

January 22, 2020

**Abstract**

Online misinformation creates serious challenges for policy-makers. The challenge is especially acute in developing countries such as India, where misinformation tends to be disseminated on private, encrypted discussion apps such as *WhatsApp*. Faced with this challenge, platforms have encouraged users to correct the misinformed beliefs they encounter on the platform. When, if at all, should we expect this user-driven strategy to be effective? Specifically, to what extent does the source and the sophistication of the corrective messages posted by users affect their ability to dispel misinformation? To explore these questions, we experimentally evaluate the effect of different types of corrective on-platform messages on the persistence of seven common rumors, among a large sample of social media users in India (N=5104). In keeping with existing results on corrections on other platforms and/or in other contexts, we show that corrections can be effective, albeit on *some* rumors and not others. More importantly, we show that the source and the sophistication of corrective messages does *not* strongly condition their effect: brief and unsourced corrections achieve an effect comparable to that of corrections implying the existence of a fact-checking report by a variety of "credible" sources (domain experts and specialized fact-checkers alike). This suggests that merely signaling a doubt about a claim (regardless of *how* detailed this signal is) may go a long way in reducing misinformation. This has implications for both users and platforms.

**Keywords**: Misinformation; Social Media; Correction; Motivated Reasoning; *WhatsApp*.

*Thanks to Ipsa Arora, Ritubhan Gautam, and Hanmant Wanole for research assistance. Funding is from Facebook. The authors thank Alex Leavitt and Devra Moelher at Facebook, as well as participants to the Facebook Integrity Research Workshop (June 2019).

# 1 Introduction

Contemporary social media platforms offer a rich ground for the spread of misinformation - false or inaccurate information such as rumors, insults and pranks. Since the 2016 US election, there has been widespread concern that misinformation on social media distorts public opinions, reduces trust in democracy or even encourages violence.

While the vast majority of the literature on misinformation has so far focused on the United States, the crisis is global. Research suggests 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). Specific studies suggest that social media misinformation played a major role in the recent presidential election in Brazil; that misleading information was used to incite violence against the Rohingya in Myanmar; or that minorities are frequently targeted in Sri Lanka because of online rumors.[1] India - our empirical focus in this study - has over the past few years emerged as a hotspot in this global misinformation crisis. In recent years, messages containing misinformation about electoral security, terrorism, HIV/AIDS, and vaccine safety have circulated widely in the country, arguably leading to a slay of problematic behaviors offline (murders, riots, and other harmful behaviors), and possibly affecting the outcome of elections (Sinha et al., 2019).

In many of these cases, misinformation is disseminated through encrypted discussion apps, the most common of which is WhatsApp. According to the company, 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. The expansion of digital economy in

[1] 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

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 (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, India is WhatsApp's biggest market in the world, with over 400 million users connected to the app, reaffirming its popularity and gigantic reach.[2].

Further, 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. This encryption calls for a different type of intervention than misinformation on other platforms. A salient one, recently promoted by WhatsApp itself, is to encourage users to fact-check information and to correct misperceptions they encounter on their private discussion threads - a strategy we refer to as "user-driven correction".

When, if at all, should we expect such a strategy to be effective? Specifically, to what extent does the source and the sophistication of the corrective messages posted by users affect one's ability to counter misinformation? To explore these questions, we experimentally evaluate the effect of different types of corrective, user-driven messages on the persistence of misinformed beliefs among a large sample of social media users in India (N=5104). This study constitutes one the first attempts to evaluate users' percep-

[2]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-report/articleshow/68288868.cms

tions of misinformation and corrective messages on an encrypted discussion app, on which users by design control input. It is also one of only a handful of studies to focus on this question *in India*. This allows us to explore the extent to which social media misinformation and potential corrections are perceived in a context of low literacy and low digital literacy, and in which partisan identities may play a weaker role than in the US.

In keeping with existing findings on fact-checking on other social media platforms and/or in other contexts, we show that user-driven corrections can be effective relative to a no correction condition, albeit on *some* rumors and not others. We also show that the heterogeneous effects partisanship and motivated reasoning are negligible, pointing at important differences in the mechanisms through which misinformation persists across contexts (Nyhan and Reifler 2011). More importantly, we show that the source and the sophistication of corrective messages does *not* strongly condition their effect: in our experiment, brief and unsourced corrections achieve an effect comparable to that of corrections implying the existence of a fact-checking report by a variety of "credible" sources (domain experts and specialized fact-checkers alike). This suggests that corrective messages may need to be frequent rather than sourced or sophisticated, and that merely signaling a problem with the credibility of a claim (regardless of *how* detailed this signaling is) may go a long way in reducing overall rates of misinformation.

This has implications for both users and platforms. Rather than unrealistically expecting users to refer to fact-checking reports (a practice they are unlikely to engage in in the first place), users should be encouraged to effectively "sound off" as easily as possible and express their doubts about on-platform claims. The group-based nature of chat applications such as WhatsApp may ensure that social pressures to not give in to misinformation decrease expressed beliefs in fake news. Our results suggest that creating a simple "button" to express doubt in reference to on-platform claims may be a complementary, cost-effective way to limit rates of beliefs in common online rumors.

# 2 The Next Frontier of Fact-checking: WhatsApp in Developing Countries.

For the past few years, platforms and policy makers have deployed a variety of strategies to combat misinformation on social media. While public platforms such as Facebook have taken to algorithmic changes and to suppressing problematic content (Nyhan et al 2017, Clayton et al 2019, Bode and Vraga 2015, Pennycook et al 2017), no equivalent solution exists for discussion apps, where encryption has been, and is likely to remain, central to the branding of the service.

As a result, platforms such as *WhatsApp* have turned to an array of alternative strategies. *WhatsApp* has encouraged media literacy education and training; short of being able (or willing) to control content, changes to the interface (limits to the number of forwards) have also been implemented. But the public outcry over violent incidents linked to rumors disseminated through the platform has also led the platform to encourage user-driven fact-checking. *WhatsApp* has bought full-page ads in multiple Indian dailies ahead of the 2019 elections (Appendix 1), asking users to seize upon fact-checked information in order to potentially correct the claims made by other users. To what extent should we expect such a strategy - so far the only known strategy to correct misinformation on encrypted discussion apps - to be effective?

## 2.1 What Do We Know About Fact-Checking?

The effectiveness of *user-driven* fact-checking depends, more generally, on the effectiveness of fact-checking. In order to hypothesize about the effect of user-driven corrections, it is thus useful to review the existing evidence about fact-checking.

A vast research agenda has over the past decade explored the effect of 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, this research has however been met with mixed success: fact-checking and warning treatments are only effective when misinformation is not salient, when priors are weak, and when outcomes are measured immediately after an intervention. Besides, 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).

The dominant theoretical explanation for this persistence comes from research on motivated reasoning (Flynn et al., 2017). According to Kunda (1990), when people process information different goals may be activated, including directional goals (trying to reach a desired conclusion) and accuracy goals (trying to process the most correct form of the information). In the context of political misperceptions, the term motivated reasoning typically refers to directionally motivated reasoning, leading people to seek out information that reinforces their preferences (confirmation bias), counter-argue information that contradicts their 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). Further, directional motivations are responsible for the stability of misperceptions over time (Nyhan, 2012). Finally, directional motivations may exacerbate the continued influence of false information even after it has been debunked (Bullock 2007; Thorson 2015a).

In sum, motivated reasoning limits the ability of corrective interventions to succeed. 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, most likely because of motivated reasoning.

To what extent should we expect these findings to apply to discussion apps and to India? In the rest of this section, we detail why corrections disseminated *on WhatsApp* (by design, user-driven corrections) and corrections in developing countries such as India - where *WhatsApp* is prevalent - may not follow these patterns.

## 2.2 Beyond Facebook, Beyond the US: Fact-checking *Discussion Apps* in *Developing Countries*.

In order to think through what the likely impact of user-driven fact-checking on chat apps might be, it is useful to consider how the contexts in which the service is most frequently used might affect the efficiency of corrections, as well as the way in which the *user-driven* nature may affect the same outcome.

Starting with the context, it is obvious that social media users in developing countries - where *WhatsApp* is most popular - may be different, on a number of crucial dimensions. India is a case in point. India has a relatively low 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%, relative to Kenya (78%), Sri Lanka (93%), Myanmar (93%), and Brazil (93%). 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. These low education and literacy rates likely 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). But they also may make corrections more difficult - as readers may not as easily understand the content or the motivation of a detailed correction. Hence we may expect a higher vulnerability on average to misinformation and misperceptions among populations with lower literacy and lower education, as well as a smaller tendency to change belief after exposure to a correction.

Relatively low rates of *digital* literacy may in turn imply that corrections would affect beliefs differently. 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. Low digital literacy likely implies that news received via the internet might automatically have more value given the unfamiliarity and fascination the medium inspires. While this may lead misinformation to be more easily believed, the same may apply to corrections. Following this line of reasoning, we may expect relatively unsophisticated corrections to have a beneficial effect.

Populations newly connected to the internet may also experience social media differently. 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.

The relative weakness of partisanship in most emerging democracies (Mainwaring and Zoco 2007) may in addition imply that motivated reasoning would *not* constitute as big an obstacle to correcting beliefs. Chibber and Verma (2018) describe how contemporary politics in India is traditionally viewed as chaotic, volatile and non-ideological in nature. Indian politicians have repeatedly 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 dictate political attitudes. Each of these elements casts some doubt on whether partisan motivated reasoning operates in the same way that it does in the American context.

Beyond these contextual factors, it is necessary to consider the *user-driven* nature of fact-checking on *WhatsApp* in order to form expectations as to its effect. This likely constitutes a challenge for fact-checking, for at least three reasons. First, because corrections are by design likely to remain short, in light of the burden that typing long and detailed corrective messages places on users. Second, discussion apps corrections can only - by design - be considered intermediated corrections: that is, even if users formally cite a source in their short corrective message (which we cannot always expect them to do), it is likely the case they will also personally be seen as one of the sources of said correction. Insofar as participants are most frequently likely to not be seen as

authoritative sources for corrections, one may doubt the efficiency of user-driver corrections. This may however be attenuated by naturally high levels of homophily in the composition of *WhatsApp* groups. Third and finally, user-driven fact-checking opens the door to a variety of formats, sources and levels of sophistication in corrective messages. Users-posted corrective messages may look extremely different, on several dimensions. Corrective messages may be short or long; sourced or not; and if sourced, they may originate from a wide variety of sources. In India, following a trend started in the United States in the mid 2010s, a host of organizations and institutions have over the past three years launched initiatives to correct online misinformation. Actors connected to the state (officials, policemen, teachers), domain experts, journalists, fact-checking organizations and social media platforms themselves have all joined on this effort. Which of these actors potential fact-checking users end up relying on may affect the effectiveness of their correction, and beyond, their ability to dispel misinformation.

## 2.3 Hypotheses

when, if at all, is the strategy of user-driven corrections to misinformation effective? Specifically, to what extent does the source and the sophistication of the corrective messages affect their ability to dispel misinformation?

To answer these questions, we evaluate the effect that different user-posted corrective messages have on belief in common rumors. We vary both the source cited by the correcting user and the degree of sophistication or length of her corrective message. As detailed below, we test for the effect of no fewer than seven different sources of corrections, representative of the diversity of actors who engage in fact-checking in India, compared to a control (no correction) and pure control (rumor is not presented in *WhatsApp* format, nor is it corrected). Given our general lack of clear priors as to how these different types of correctives might compare in the Indian context, and given the contradictory intuitions listed in the precedent section, we did not lay down in our pre-

analysis plan a formal hypothesis about this central question (instead labeling it as a "research question"). Despite this, our primary interest remains to compare whether different corrective messages differently affect belief rates.

Since we are so far lacking evidence about the effect of corrections *on WhatsApp threads* and *in India*, we also pool across these different types of corrections, which allows us to test the following hypothesis:

H1: Exposure to corrective information on *WhatsApp* (pooling together the different types of correction) will reduce the perceived accuracy of the targeted claim.

As noted above, the literature on corrections has so far identified motivated reasoning as a significant hurdle for corrections. In addition to exploring the effect of different types of user-driven corrections on beliefs in rumors, we thus test whether potential corrective effects are conditional on the type of rumors or on the perceived ideological bias of its source(s). In keeping with leading hypotheses in the literature on fact-checking, we also investigate the potential interaction of correction with the perceived partisanship or rumors and of their sources; while studies set in the American contest have repeatedly shown the influence of motivated reasoning on information processing (Taber and Lodge 2006; Nyhan and Reifler 2012), no such evidence exists in contexts such as India in which partisan affiliations may be less strong or stable in the first place. We thus explore the following hypotheses:

H2a: Users' corrections will be more effective when the targeted claim is attributed to a dissonant politician (compared to when it is unattributed).

H2b: Users' corrections will be less effective when the targeted claim is attributed to a congenial politician (compared to when it is unattributed).

H3a: Users' corrections will be more effective when the targeted claim originates from a dissonant media outlet (compared to an unattributed or neutral outlet).

H3b: Users' corrections will be less effective when the targeted claim originates

from a congenial media outlet (compared to an unattributed or neutral outlet).

H4a: Users' corrections will be less effective when the targeted claim is ideologically congenial to respondents (compared to non-ideological claims).

H4b: Users' corrections will be more effective when the targeted claim is ideologically dissonant to respondents (compared to non-ideological claims).

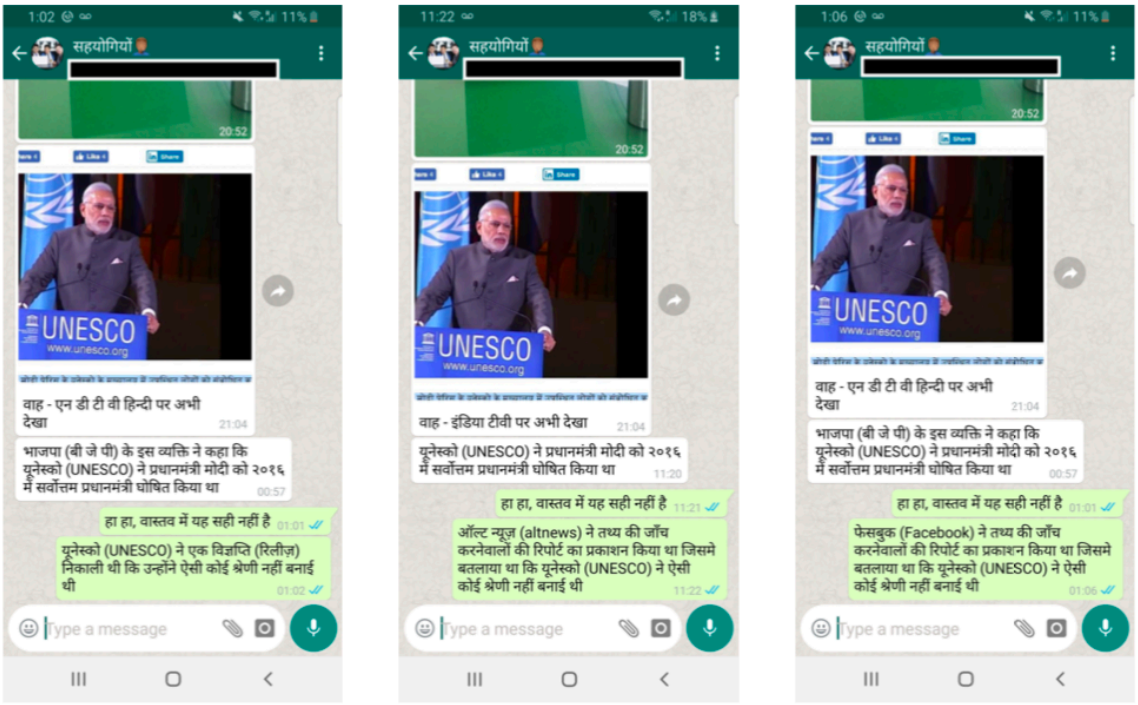
# 3 Design

We test these hypotheses with a survey experiment conducted on a large sample of Hindi speakers (N=5104) recruited from *Facebook*[3], a common platform where misinformation is shared in India. Our experiment examines the effectiveness of corrective messages included on a fictitious but realistic *WhatsApp* chat screenshot. We show respondents recruited on *Facebook* screenshots of a credible *WhatsApp* thread featuring a controversial claim (note that respondents are themselves not a part of the chat, they merely see a fictitious screenshot). The screenshot is one of a group-chat in the WhatsApp application, where two members of the group are shown to be discussing a rumor. Our experiment examines the effect of including various types of corrective messages by a participant in this fictitious chat on the respondent's rate of belief in the claim made by the first participant.

Figure 1 below provides a visual example of the screenshots different respondents were exposed to in the context of one of the nine claims we experimented with (below referred to as "UNESCO claim"): as can be seen here, the last message provides slightly different versions of a corrective message. While we manipulate various other aspects of the thread, the effect of variations in this last corrective message are the main focus of our empirical analyses below.

[3]The Facebook ad used to recruit respondents is presented in Appendix 2. Eligibility and exclusion criteria for participants: Hindi speaking Facebook users located in India (age 18+).

Figure 1: Different versions of a prompt.



The survey that respondents are recruited for through Facebook consists of two sections: a pre-treatment section and an experimental section. The pre-treatment section includes questions about respondents' demographics and pre-treatment covariates: political attitudes, conspiratorial predispositions, and trust in various political and health institutions.[4]

The experimental section includes our experimental treatments and outcome measures. The experimental prompts require respondents to read and evaluate *WhatsApp* chat screenshots about political, health, and social claims that have circulated recently in India. Each thread focuses on one factual claim that we subsequently ask respondents to evaluate. In order to ensure respondents were used to the format of the prompts and build up the expectation that some of the claims included in the experiment were "true", all respondents began this section by rating the perceived accuracy of a "true claim" (that is, one featuring a verifiable and uncontroversial claim): that Australia holds the record number of victories in the cricket world cup (see below), a fact widely known in the sampled population. The order of the remaining 8 claims was randomized. In total, all respondents evaluated 9 claims[5], 7 of which are false (F) and 2 of which are true (T), including the aforementioned Australia claim. These claims are:

1. "Australia is the country that has won the ICC cricket world cup the most often" (T) – NOTE: respondents were always exposed to this "true" claim first.

2. "There is no cure for HIV/AIDS" (T).

3. "In the future, the Muslim population in India will overtake the Hindu population in India" (F).

4. "Polygamy is very common in the Muslim population" (F).

[4]We explore the correlation between these variables and rates of beliefs in rumors and corrective messages in a related paper.

[5]unless of course there was attrition, which was the case for less than 1% of our sample.

5. "M-R vaccines are associated with autism and retardation" (F).

6. "Drinking cow urine (gomutra) can help build one's immune system" (F).

7. "Netaji Bose did NOT die in a plane crash in 1945" (F).

8. "The BJP has hacked electronic voting machines" (F).

9. "UNESCO declared PM Modi best Prime Minister in 2016" (F).

These 8 claims were chosen following a pretest during which we evaluated baseline levels of beliefs in a large series of claims. Each of these claims were believed or strongly believed by at least of 25% of the population of the pretest, with some of these statements being believed by a large *majority* of respondents. Data from this pretest is presented in the Appendix.

The final selection of claims was the product of several constraints and choices. Qualitative intuitions and pretesting first led us to determine that 9 claims was the maximum number of claims we could expose respondents to without expecting major decreases in attention or increases in attrition. We subsequently determined that we needed a few of the claims to to be "true" in order to avoid prompting respondents to systematically reject the veracity of claims. But simultaneously, we wanted to maximise respondent exposure to controversial fake political claims that were spread widely during the run up to the 2019 elections in India. We chose to select claims touching on a variety of topics, in order to build on the current literature on misinformation and corrections. Specifically, we decided to select claims about 4 prevalent topics: current electoral politics (claims 8 and 9), health (claims 5 and 6), minorities (3 and 4) and historical conspiracies (claim 7). Finally, of the claims for which we experimentally planned to manipulate the presumed identity of the speaker (claims 3,4,6,8,9), we wanted to select claims that could have been credibly made by a diversity of partisan actors. As shown by the results of the 2019 elections, one party (the BJP) currently dominates the debates

and the informational environment in India; as a result, many more claims would be expected to be made by BJP politicians than by other, opposition politicians. Many of the prevalent misinformation we select from as part of the pretest thus naturally tends to be associated with the BJP. To reflect this, we mostly include claims whose speaker could only credibly be a BJP politician (claims 3,4,6, and 9). But we make sure (by including claim 8) that we have at least one prevalent claim that could credibly be made by an opposition politician - specifically, in our case, an INC (Indian National Congress) politician.

While it might have been equally interesting to include alternative claims that we pretested, we believe this selection to be somewhat representative - both in its thematic focus and in the diversity of themes touched upon - of the type of misinformation that is commonly peddled on Indian social media. As a point of reference, the exhaustive referencing of misinformation presented in Sinha et al (2019) presents a sample of claims relatively comparable to ours.

For each claim, respondents were randomized into one of several treatment conditions or, for a very small percentage of them (3%), to a "pure control condition" in which they were not exposed to any thread, but instead simply asked whether they believed in one of the nine claims listed above.[6] Respondents who were assigned to read a thread (all except those assigned to the pure control) had equal probability of being assigned to each of the possible combinations of experimental treatments listed below.

To increase realism, we excluded highly unrealistic manipulations (e.g., voter fraud allegations attributed to ruling party politicians) and tailored domain expert corrections to each rumor (e.g., we attribute expert corrections of voter fraud rumors to the Election Commission of India).

Given that respondents each saw nine threads, we had to vary the text of the mes-

[6] As shown in Appendix 11, we detect no differences in the overall rate of belief in all 7 "false" claims when comparing our pure control condition (no thread is shown; respondents are only asked to evaluate the veracity of the claim) to our control condition (WhatsApp thread is shown, but not corrected, after which respondents are asked to evaluate the veracity of the claim).

sages in order to ensure realism. The spreadsheet presented in Appendix 3 gives a full list of treatments for each rumor used in our experiment. When detailing our treatment below, we also refer to this sheet. But the principle was the same across the nine threads. The first participant to the thread in each case posted a visual of a press article (in message 1) and described (in message 2) the content of the article - potentially identifying in the process the publication on which she/he found the claim and the politician that had made the claim. Subsequently, a second participant reacted to these posts, in some cases attempting to correct participant 1's claim (through a diversity of strategies), in other cases simply thanking participant 1 for her/his contribution. In order to avoid biasing responses, we deliberately blackened the purported names of the participants and presented this as a measure to protect their privacy - hence likely increasing the realism of the experiment.

Variations on each of the three experimental factors were as such:

**Experimental factor 1**: media outlet reporting false claim - 3 possible values:

1. NDTV Hindi (anti-BJP private channel).
2. India TV (clearly pro-BJP private channel).
3. Doordarshan News (public channel - hence presumably pro-government and pro-BJP).

**Experimental factor 2**: identity of the politician making false claim - 3 possible values overall (but only 2 on each claim; either 1 and 3 or 2 and 3):

1. BJP party politician.
2. INC party (anti-BJP) politician.
3. No attribution. In this case, the text simply relies on a passive voice ("it has been said that...") or a vague attribution ("some have said...").

**Experimental factor 3**: presence/source/sophistication of corrective message on the displayed thread. Respondents were equally likely to be assigned to one of four possible conditions here, though the fourth condition was further subdivided (equally) in five subcategories, as detailed below.

1. no correction (referred above as "control" condition). In this case the second participant simply thanked the first participant for posting and/or said they would have a look at the article posted (full text for each thread in Appendix 3).

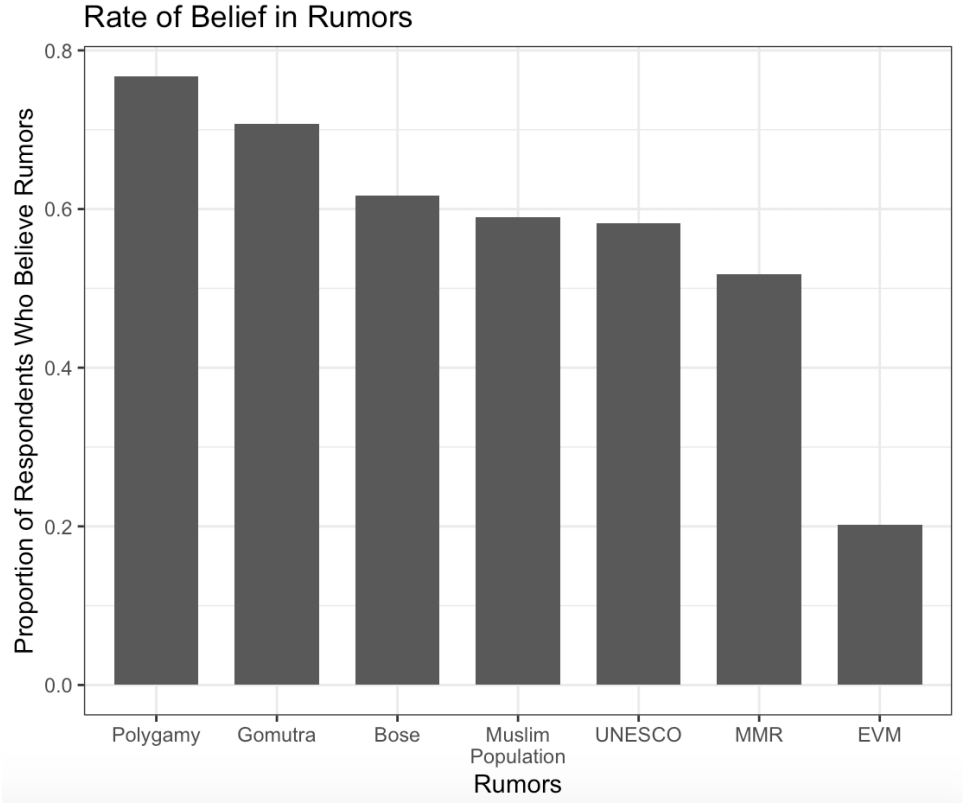
2. "random guy" correction. This consisted in a short and unsourced non-expert correction. As shown in Appendix 3, participant 2 in this case simply voices their incredulity, but did not cite a reason or a source, making the message shorter in the process.

3. Authority/domain expert correction. In this case, the second participant attemopts to correct the claim by citing an authority relevant to the claim being discussed, for instance a medical professional re. Health claims, or the electoral commission re. EVMs.

4. Correction by a specialized fact-checking service, such as:

    - Altnews – in this case, the second participant argues in his correction that this specialized fact-checking organization with left-leaning politics has fact-checked the claim and found it to be erroneous.
    - Vishwasnews - in this case, the second participant argues in his correction that this specialized fact-checking organization with right-leaning politics has fact-checked the claim and found it to be erroneous.
    - The Times of India – in this case, the second participant argues the country's best known/oldest newspaper has fact-checked the claim and found it to be erroneous.

**Figure 2: Overall rates of belief in rumors across experimental conditions**




- Facebook – in this case, respondents are told that the online platform itself (Facebook) has fact-checked the claim and found it to be erroneous.
- WhatsApp – in this case, respondents are told that the online platform itself (WhatsApp) has fact-checked the claim and found it to be erroneous.

After reading each of the threads, respondents answered a single outcome question about their belief in the claim:

*How accurate is the following statement?* [Statement of the claim]

(very accurate, somewhat accurate, not very accurate, not at all accurate)

# 4 Results

Across all conditions, and despite the presence of a corrective message on 3/4 of our prompts, respondents were remarkably likely to declare that the false claims made by the first participant were either accurate or somewhat accurate. As shown in Figure 2, 6 of the 7 false claims were rated as accurate or somewhat accurate by at least half of our sample. Only the claim about EVMs was rated as accurate by a minority of our sample, which seems in line with the fact that our sample was heavily pro-BJP (see descriptives in Appendix).

How did the presence of different types of corrections and other manipulations affect these average rates of belief?

## 4.1 H1: Do Corrective Messages Matter?

We first test H1. As a reminder, H1 states that "exposure to corrective messages on WhatsApp will reduce the perceived accuracy of the targeted claim".

To test H1, we pool together all the different types of user-driven corrections and estimate a separate model for each claim listed above. In Table 1, we simply show bivariate OLS models in which we evaluate the impact of being exposed to *any* type of user-driven correction (as opposed to being exposed to the control condition) on our dependant variable.[7] In table 2, we simply add controls for all other experimental manipulations included in our design, which unsurprisingly does not change our estimates of the effect of being exposed to a corrective message. Figure 3 simply provides a graphical summary of Table 1 and maps how user-driven corrections impact our dependant variable.

As is clear from Figure 3, respondents *do* react to the addition of a corrective mes-

[7]Note that we omit the pure control from these analyses. Clubbing them with the control condition does not however change our results, as there is overall no difference between the control and pure control conditions (Appendix 11).

sage. Exposure to user-driven fact-checking appears to reduce the likelihood that they report the claim to be accurate or somewhat accurate. The only claim for which we do not identify a statistically significant result is the EVM claim, which appears to be believed by a much smaller pool of respondents in the first place, making it much harder for us to detect an effect of this type. The picture is however far from consistent across the other claims, as effect sizes are rather diverse. While the corrective effect on most claims is small (and close to insignificance), it is remarkably large (above 0.4 on a scale from 1 to 4) on two of the items: the MMR vaccine claim and the UNESCO claim.

That such an effect exists may be seen as good news for the platform, insofar as it suggests that the user-driven strategy advocated by *WhatsApp* may reduce overall rates of beliefs in patently false claims circulating on threads. It is however necessary to remain cautious, for several reasons. The first one owes to the experimental context in which we are able to detect these effects. While this is difficult to evaluate, it is not clear whether the dosage of these corrections is perfectly calibrated to reproduce a real-world experience. On the one hand, it is possible that participants to the experiment were on average exposed more intensely to these corrections than they would be in the real world. On the other hand, it is quite remarkable that they reacted to a correction posted by an unidentified individual posting on a thread they are themselves not part of. Nothing clearly incentivized them to do so, and it is credible that corrections posted by a well-known individual on an actual thread built on homophily would be *more* impactful. The second reason to remain cautious owes to the variation across threads. Figure 3 broadly suggests that respondents were more open to being corrected on some claims than on others, in a way that cannot be easily explained. Importantly, the size of the effect across rumors does not seem related to the prior salience of these rumors in our sampled population. As shown in Appendix 10, many respondents in our pretest had heard of the widely circulated UNESCO rumor, while much fewer had heard of the claim about MMR vaccines. Yet both led to comparatively large corrective effects.

Table 1: Main Effect of Corrections on Belief in Rumors

| | *Dependent variable: Belief in Rumor* | | | | | | |
|---|---|---|---|---|---|---|---|
| | MuslimPop | Polygamy | MMR | Gomutra | EVM | UNESCO | Bose |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Correction | −0.104*** | −0.190*** | −0.448*** | −0.106*** | −0.024 | −0.411*** | −0.109*** |
| | (0.037) | (0.030) | (0.033) | (0.033) | (0.032) | (0.040) | (0.031) |
| Constant | 2.746*** | 3.272*** | 2.827*** | 3.010*** | 1.650*** | 2.999*** | 2.788*** |
| | (0.033) | (0.026) | (0.028) | (0.027) | (0.027) | (0.034) | (0.025) |
| Observations | 5,104 | 5,103 | 5,061 | 5,099 | 5,136 | 5,109 | 5,117 |
| $R^2$ | 0.002 | 0.008 | 0.035 | 0.002 | 0.0001 | 0.021 | 0.002 |
| Adjusted $R^2$ | 0.001 | 0.008 | 0.035 | 0.002 | −0.0001 | 0.020 | 0.002 |
| Res. Std. Er. | 1.095 | 0.946 | 1.039 | 1.060 | 1.014 | 1.251 | 1.048 |
| F Statistic | 7.859*** | 39.895*** | 185.869*** | 10.536*** | 0.568 | 107.789*** | 12.390*** |

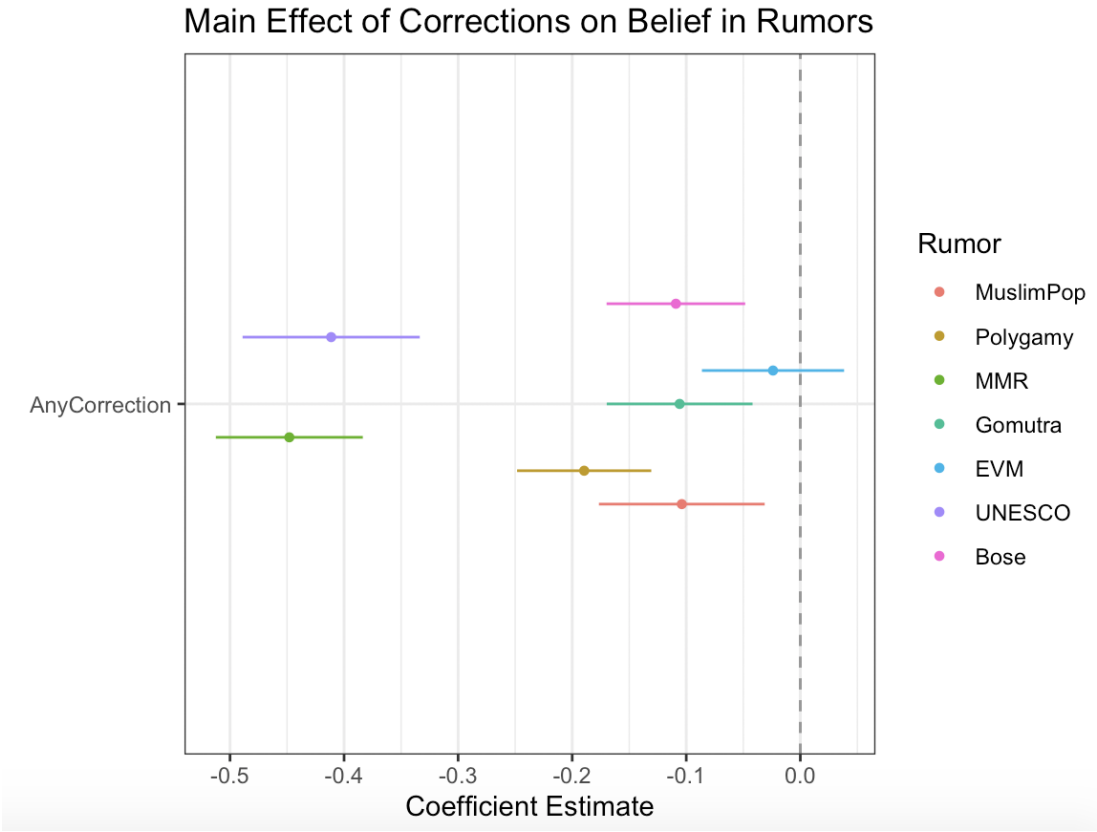
*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 2: Main Effect With Controls

| | *Dependent variable: Belief in Rumor* | | | | | | |
|---|---|---|---|---|---|---|---|
| | MuslimPop | Polygamy | MMR | Gomutra | EVM | UNESCO | Bose |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Any Correction | −0.106*** | −0.182*** | −0.428*** | −0.112*** | −0.018 | −0.413*** | −0.116*** |
| | (0.037) | (0.031) | (0.033) | (0.033) | (0.032) | (0.040) | (0.032) |
| Dissonant Media | 0.084* | −0.011 | −0.107** | 0.059 | −0.087* | 0.016 | 0.030 |
| | (0.050) | (0.044) | (0.046) | (0.049) | (0.046) | (0.059) | (0.047) |
| Congenial Media | 0.041 | 0.101** | −0.119*** | 0.048 | −0.119** | −0.020 | 0.037 |
| | (0.049) | (0.045) | (0.045) | (0.048) | (0.046) | (0.059) | (0.045) |
| Copartisan Speaker | −0.022 | −0.008 | | 0.091* | 0.434*** | 0.091 | |
| | (0.051) | (0.046) | | (0.050) | (0.061) | (0.060) | |
| Outpartisan Speaker | −0.190*** | −0.271*** | | −0.297*** | −0.141*** | −0.157** | |
| | (0.064) | (0.057) | | (0.065) | (0.047) | (0.080) | |
| Constant | 2.747*** | 3.275*** | 2.840*** | 3.008*** | 1.662*** | 2.999*** | 2.785*** |
| | (0.033) | (0.026) | (0.028) | (0.028) | (0.027) | (0.034) | (0.025) |
| Observations | 5,104 | 5,103 | 5,061 | 5,099 | 5,136 | 5,109 | 5,117 |
| $R^2$ | 0.004 | 0.013 | 0.037 | 0.008 | 0.015 | 0.022 | 0.003 |
| Adjusted $R^2$ | 0.003 | 0.012 | 0.037 | 0.007 | 0.014 | 0.021 | 0.002 |
| Res. Std. Er. | 1.094 | 0.943 | 1.038 | 1.057 | 1.007 | 1.250 | 1.048 |
| F Statistic | 3.599*** | 13.555*** | 65.656*** | 8.161*** | 15.753*** | 23.216*** | 4.443*** |

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

**Figure 3: The Effect of User-driven Corrections Across Rumors**




While additional research is necessary to reach a conclusion on this point with a higher degree of certainty, it is possible that claims fitting in a core belief system (either religious or partisan, or mixed as in the case of the Hindu BJP supporters that comprise a majority of our sample) may be harder to correct.

## 4.2 How Effective Are Different Types of User-driven Corrections?

Moving beyond the overall effect across types of corrections, our next objective is to decompose this effect. This allows us to observe whether some variations of corrective messages are more effective than others.

This is what we do in Table 3 and Figure 4. In Table 3, we rely on OLS models to evaluate the effect of different types of corrections on the level of belief in each of the claims, compared to the control condition (omitted category).[8] Figure 3 provides a

[8]We similarly exclude the pure control from these analyses; Note however that it does not change our

graphical representation of this regression table.

Several interesting findings emerge from these results. Most importantly, we do not observe dramatically different effect sizes across sub-types of corrections: confidence intervals between any two of these corrections, on any of these rumors, overlap. This broadly implies that the length and the sophistication of user-driven corrections matters very little. As can be seen from Figure 4, the "random guy" correction is often as effective as the longer and more clearly sourced corrections we experiment with in this design. An unidentified participant merely expressing incredulity about a claim is thus as likely to reduce belief in a falsehood as a more carefully crafted - but maybe more unlikely - correction. Even more surprising, reference in the corrective message to a domain expert does not appear to make the correction more persuasive: in all cases, respondents are as likely to react to the correction when it is said to originate from a professional fact-checking organization, a prominent newspaper (TOI) or the platforms themselves (bottom two rows), as opposed to a domain expert. This further implies that respondents open to belief change do not require much "expertise" in order to be moved, which further reinforces the inference we can draw from the estimate on the "random guy correction" experimental group. Finally, and more generally, no source emerges as consistently more persuasive or effective from this exercise. While the differences are far from significant, Facebook emerges as more effective than WhatsApp, and the platforms appear as effective as the fact-checking specialists whose work they encourage. This may imply that outsourcing fact-checking may not be necessary to improve its overall credibility. All of this more generally proves the above point: that the content of user-driven corrections is generally unimportant, or at least less important than the mere existence of such a correction.

results whatsoever.

Table 3: Effect of Correction Source

| | *Dependent variable: Belief in Rumor* | | | | | | |
|---|---|---|---|---|---|---|---|
| | MuslimPop | Polygamy | MMR | Gomutra | EVM | UNESCO | Bose |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Peer ("Random guy") | $-0.076^{*}$ | $-0.126^{***}$ | $-0.348^{***}$ | $-0.101^{**}$ | $-0.013$ | $-0.292^{***}$ | $-0.091^{**}$ |
| | (0.045) | (0.037) | (0.041) | (0.042) | (0.039) | (0.049) | (0.039) |
| Expert | $-0.119^{***}$ | $-0.234^{***}$ | $-0.469^{***}$ | $-0.149^{***}$ | $-0.016$ | $-0.483^{***}$ | $-0.125^{***}$ |
| | (0.044) | (0.037) | (0.041) | (0.040) | (0.041) | (0.049) | (0.043) |
| AltNews | $-0.079$ | $-0.199^{***}$ | $-0.591^{***}$ | $-0.092$ | $-0.040$ | $-0.487^{***}$ | $-0.171^{**}$ |
| | (0.075) | (0.067) | (0.069) | (0.069) | (0.068) | (0.086) | (0.072) |
| Vishwas | $-0.134^{*}$ | $-0.276^{***}$ | $-0.468^{***}$ | $-0.101$ | $-0.038$ | $-0.367^{***}$ | $-0.075$ |
| | (0.074) | (0.064) | (0.069) | (0.071) | (0.067) | (0.090) | (0.074) |
| TOI | $-0.115$ | $-0.248^{***}$ | $-0.522^{***}$ | $-0.046$ | $-0.051$ | $-0.398^{***}$ | $-0.207^{***}$ |
| | (0.077) | (0.065) | (0.071) | (0.073) | (0.068) | (0.085) | (0.072) |
| Facebook | $-0.186^{**}$ | $-0.160^{**}$ | $-0.529^{***}$ | $-0.115$ | $-0.040$ | $-0.547^{***}$ | $-0.102$ |
| | (0.074) | (0.066) | (0.073) | (0.070) | (0.066) | (0.088) | (0.072) |
| WhatsApp | $-0.070$ | $-0.158^{**}$ | $-0.494^{***}$ | 0.034 | $-0.040$ | $-0.529^{***}$ | $-0.028$ |
| | (0.073) | (0.065) | (0.072) | (0.074) | (0.070) | (0.091) | (0.071) |
| Constant | $2.746^{***}$ | $3.272^{***}$ | $2.827^{***}$ | $3.010^{***}$ | $1.650^{***}$ | $2.999^{***}$ | $2.788^{***}$ |
| | (0.033) | (0.026) | (0.028) | (0.027) | (0.027) | (0.034) | (0.025) |
| Observations | 5,104 | 5,103 | 5,061 | 5,099 | 5,136 | 5,109 | 5,117 |
| $R^2$ | 0.002 | 0.010 | 0.039 | 0.004 | 0.0002 | 0.025 | 0.003 |
| Adjusted $R^2$ | 0.001 | 0.009 | 0.038 | 0.002 | $-0.001$ | 0.024 | 0.002 |
| Res. Std. Er. | 1.095 | 0.945 | 1.037 | 1.060 | 1.014 | 1.249 | 1.048 |
| F Statistic | 1.589 | $7.424^{***}$ | $29.487^{***}$ | $2.585^{**}$ | 0.174 | $18.591^{***}$ | $2.520^{**}$ |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

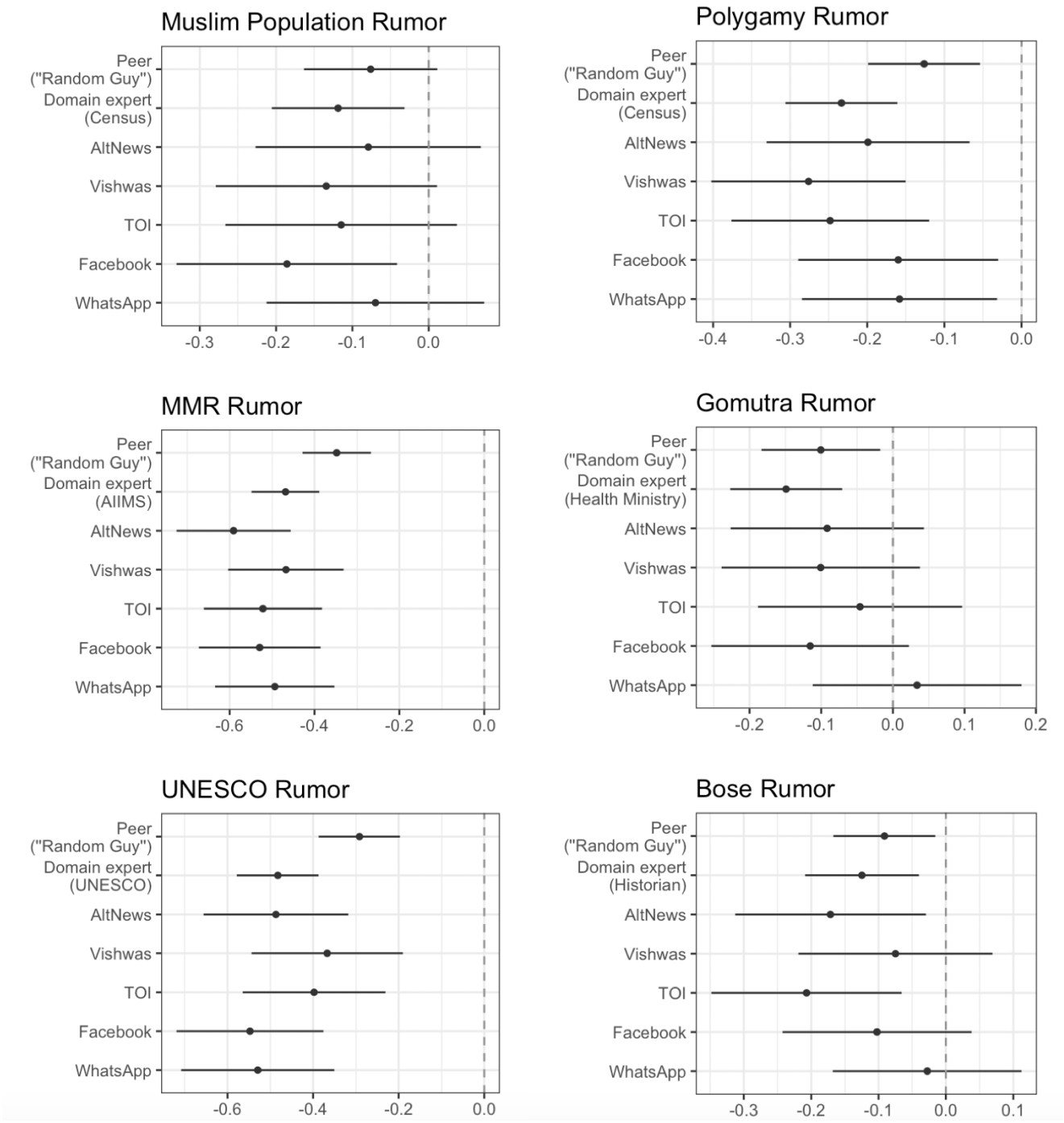



Figure 1: Effect of Correction Source

## 4.3 Motivated Reasoning and Corrections

To what extent are the limited effects detected above a consequence of motivated reasoning? Specifically, to what extent are these effects conditional on the party identity of the political actor to whom the claim is attributed in the first message (H2a and H2b)? To what extent are they conditional on the media outlet on which the claim is said to have been made (H3a and H3b)? Finally, to what extent are they conditional on congeniality of the claim itself (H4a and H4b)?

To test H2a and H2b, we limit our analyses to the subset of rumors that are clearly partisan in nature (claims 3, 4, 6, 8,and 9) and code whether the claim was attributed in the prompt to a congenial or dissonant politician. We code a politician as congenial or dissonant as a function of the respondent's partisan inclination towards the BJP (the ruling party), relying on the respondent's expressed closeness to this party. Concretely, a BJP politician is deemed congenial if the respondent describes herself as close or very close to the party and dissonant if the respondent describes herself as far or very far from the party. By contrast, a INC politician is deemed congenial if the respondent describes herself as far or very far to the BJP and dissonant if the respondent describes herself as close or very close to the BJP. Note that we are pooling members of both major parties in each category (e.g., "dissonant" takes the value of 1 for BJP identifiers who read an anti-BJP claim and for INC identifiers who read a pro-BJP claim).

To test H3a and H3b, we code a media outlet as congenial or dissonant as a function of the respondent's expressed proximity to the BJP. Concretely, we code the "pro-BJP" outlet (here, India TV) as congenial and the "anti-BJP" outlet (here, NDTV) as dissonant when the respondent reports feeling close or very close to the BJP. By contrast, we code the "pro-BJP" outlet (India TV) as dissonant and "anti-BJP" outlet (NDTV) as congenial when the respondent reports feeling far or very far to the BJP.

To test H4a and H4b, we once again limit our analyses to the subset of rumors that are clearly congenial/dissonant to supporters of one of the two major national parties

Table 4: Effect of Correction * Congenial Claim on Belief in Rumor

| | *Dependent variable: Belief in Rumor* | | | | |
|---|---|---|---|---|---|
| | MuslimPop | Polygamy | Gomutra | EVM | UNESCO |
| | (1) | (2) | (3) | (4) | (5) |
| AnyCorrection | −0.092 | −0.163*** | −0.117** | 0.0002 | −0.069 |
| | (0.059) | (0.048) | (0.052) | (0.038) | (0.053) |
| CongenialClaim | 0.238*** | 0.246*** | 0.369*** | 0.520*** | 0.362*** |
| | (0.067) | (0.053) | (0.055) | (0.056) | (0.057) |
| AnyCorrection* | −0.025 | −0.043 | 0.014 | −0.057 | 0.023 |
| CongenialClaim | (0.076) | (0.061) | (0.066) | (0.066) | (0.067) |
| Constant | 2.602*** | 3.120*** | 2.784*** | 1.478*** | 2.751*** |
| | (0.052) | (0.041) | (0.043) | (0.032) | (0.045) |
| Observations | 5,104 | 5,103 | 5,099 | 5,136 | 5,099 |
| $R^2$ | 0.011 | 0.020 | 0.032 | 0.049 | 0.031 |
| Adjusted $R^2$ | 0.010 | 0.019 | 0.032 | 0.049 | 0.030 |
| Res. Std. Er. | 1.090 (df = 5100) | 0.940 (df = 5099) | 1.044 (df = 5095) | 0.989 (df = 5132) | 1.045 (df = 5095) |
| F Statistic | 18.919*** | 34.488*** | 56.388*** | 88.471*** | 53.490*** |

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

Table 5: Effect of Correction * Dissonant Claim on Belief in Rumor

| | *Dependent variable: Belief in Rumor* | | | | |
|---|---|---|---|---|---|
| | MuslimPop | Polygamy | Gomutra | EVM | UNESCO |
| | (1) | (2) | (3) | (4) | (5) |
| AnyCorrection | −0.127*** | −0.195*** | −0.088** | −0.004 | −0.032 |
| | (0.045) | (0.036) | (0.039) | (0.049) | (0.040) |
| DissonantClaim | −0.206*** | −0.188*** | −0.267*** | −0.608*** | −0.267*** |
| | (0.069) | (0.055) | (0.058) | (0.053) | (0.060) |
| AnyCorrection* | 0.062 | 0.016 | −0.060 | −0.022 | −0.058 |
| DissonantClaim | (0.078) | (0.064) | (0.069) | (0.062) | (0.070) |
| Constant | 2.816*** | 3.334*** | 3.097*** | 2.022*** | 3.058*** |
| | (0.040) | (0.031) | (0.033) | (0.042) | (0.035) |
| Observations | 5,104 | 5,103 | 5,099 | 5,136 | 5,099 |
| $R^2$ | 0.006 | 0.015 | 0.021 | 0.090 | 0.019 |
| Adjusted $R^2$ | 0.006 | 0.015 | 0.020 | 0.089 | 0.019 |
| Res. Std. Er. | 1.093 (df = 5100) | 0.942 (df = 5099) | 1.050 (df = 5095) | 0.968 (df = 5132) | 1.051 (df = 5095) |
| F Statistic | 10.658*** | 26.475*** | 36.056*** | 168.534*** | 33.051*** |

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

in India: the BJP and INC (claims 3, 4, 6, 8,and 9). We code claims as congenial or dissonant ex-ante as a function of participants own ideological inclinations and as a function of our observations of these two parties' platforms. Namely, when participants self-report being "close" or "very close" to the BJP, claims number 3 (Muslim population growth), 4 (polygamy within the Muslim population), 6 (belief about the virtues of cow urine), and 9 (Modi and Unesco) are coded as congenial claims. By contrast, claim number 8 (EVMs) is coded as dissonant, while claims number 1, 2, 5 and 7 are coded as neither congenial nor dissonant. Similarly, when participants self-report being "close" or "very close" to the INC, claim number 8 is coded as congenial while claims number 3,4,6,9 are coded as dissonant and claims 1, 2, 5, and 7 are neither congenial nor dissonant. Note that we are here again pooling members of both major parties in each category (e.g., "dissonant" takes the value of 1 for BJP identifiers who read an anti-BJP claim and for INC identifiers who read an anti-INC claim). In Appendix 7, we show that these codings are reasonable, insofar as claims a priori rated as congenial are more likely to be believed, while claims a priori rated as dissonant are less likely to be believed.

For each of these three series of hypotheses, we are interested in the potential *interaction* between exposure to a corrective message (pooling across all types of corrective messages) and the degree of congeniality of the source/media outlet/claim. Results on these three series of hypotheses point to a similar conclusion: the effect of corrections does not appear to be consistently - across rumors - conditional on the partisan positioning of the source of the claim, or on the congeniality of the claim itself. Tables 4 and 5 provide (respectively) such tests of H4a and H4b. In the interest of space tests for H2 and H3 are included in Appendix 4 and 5.

As can be seen from these results, we almost never detect a statistically significant effect on the interaction term. This suggests that partisanship and motivated reasoning (Nyhan and Reifler 2011) do not matter as much here as they frequently do in experi-

ments run on American voters, possibly pointing at important differences in the mechanisms through which misinformation persists across contexts.

This notable difference may be due to one or both of two factors we cannot easily disentangle here: the user-driven nature of the corrections presented to respondents, and the comparatively lesser partisan nature of politics in India. It is possible, first, that the correction included on our threads remains unmediated by the partisanship of the source or media on which the claim was made because participants did not identify the source of the correction himself in ideological terms, or as a member of the research team. Besides, drawing on their own experience on WhatsApp threads, participants may - correctly enough - assume that such corrections likely originate from homophilic sources, hence more easily reliable sources.

The relative weakness of partisanship in India may in addition explain why motivated reasoning does *not* appear to constitute as big an obstacle to correcting beliefs. Contemporary politics in India is traditionally viewed as chaotic, volatile and non-ideological in nature (Chibber and Verma 2018). Indian politicians have repeatedly made and unmade 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 dictate political attitudes. Each of these elements casts some doubt on whether partisan motivated reasoning operates in the same way that it does in the American context. These results may confirm that it does not: less intense forms of partisan identification may open the door for corrective effects.

Importantly, no such stable interaction exists even among the most clearly ideological and partisan subgroup in our sample: BJP supporters. In appendix 6, we run models in which we interact respondents' level of support for the ruling party and the presence of a correction on the thread. In 5 out of 7 cases, we do not detect any significant

interaction. This absence of effects persists when we run a second series of test relying on reported voting decisions in the 2019 elections instead of measuring respondents' closeness to the BJP: participants who report having voted for the ruling party in the 2019 election do not react to corrections any differently. This in our opinion confirms the lesser role played by motivated reasoning in this context.

# 5 Discussion and Conclusions

The implications of these results for the debate about misinformation on discussion apps are twofold.

These results first confirm that encouraging user-driven fact-checking *may* be beneficial. Our main analyses above overall suggest that exposure to a fact-checking message posted by an unidentified thread participant is enough to significantly reduce rates of belief in a false claim. This is important insofar as nothing incentivized participants to the experiment to pay attention to the message. Besides, they did not know and by design could not identify the individual posting this correction. It is not impossible to think that such a correction posted on a more homophilic network would achieve a much larger effect.

Expecting users to post such corrections may however be unrealistic: users may not have a good sense of what constitutes fact-checked information; they may not know of fact-checking services; if they do, they simply may not be motivated to consult these services; if they did consult them and read their analyses, they may not be willing to invest time and energy in a lengthy explanation leading them to openly contradict one of their acquaintances; even if they were willing to take these steps, they may not find the right words. In this design, we have so far assumed that such user-driven corrections may be realistic and common. But it is far from clear that they are or will become common in the real world. We simply do not have any data on this point.

Fortunately, our results suggest that they need not become common in order for platforms to effectively combat misinformation. If anything clearly emerges from our results, it is the fact that *any* expression of incredulity about a false claim posted on a thread leads to a reduction in self-reported belief. Rather than unrealistically expecting users to refer to fact-checking reports (a practice they are unlikely to engage in in the first place), users should be encouraged to effectively "sound off" as easily as possible and express their doubts about on-platform claims.

Platforms may do so in a variety of ways, and supporting fact-checking should be one of these. Since the content of corrections does not appear to matter much (as a reminder, our random guy treatment performs as well as our more sourced corrections), platforms such as WhatsApp ought to be thinking about product ideas that reduce the cost of expressing dissent on a thread. One way to do this may be to add a simple "button" to express doubt in reference to on-platform claims, or enable users to easily flag statements as problematic, unreliable, or groundless. Similar to the "like" functions that exist on other platforms, it would be technically very easy for WhatsApp to add "red flag" or "?" emoji buttons that users can easily click on next to contentious posts. Such a strategy would be entirely compatible with the encrypted nature of the platform, as "red flags" need not be reported or investigated by the platform, but merely used to communicate to other users that a variety of opinions exist among participants to the thread. Such a strategy would in addition allow a single user to very quickly flag a large number of posts, and hence more effectively combat the barrage of misinformation that currently exists on these platforms.

While these effect might be useful in countering misinformation given that it demonstrates that simple corrections in a group setting can be effective, the result also suggests that deliberately wrong "corrections" by hyper-partisan users are equally as likely to be believed. Future research should analyse whether, for instance, the "random guy" correction is effective when the random peer posts a correction that is factually in-

correct relative to accurate. Our study thus opens up broader avenues for research in countering misinformation in developing settings, and we hope to fuel a strong body of work that investigates the mechanisms for belief in user-driven corrections, the weakness of motivated reasoning in conditioning effects, as well as the demographic characteristics of corrective users in impacting beliefs.

# 6 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 per-

ceived 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.