"I Don't Think That's True, Bro!"

An Experiment on Fact-checking Misinformation in India

Sumitra Badrinathan
University of Pennsylvania

Simon Chauchard
Leiden University

D.J. Flynn
IE University*

Abstract

Misinformation on encrypted messaging applications is linked to changes in public opinion, electoral outcomes, and even violence. Since encrypted platforms cannot control content, can social media users correct their peers' misperceptions? If so, how? We experimentally evaluate the effect of different types of corrective messages on the persistence of seven common rumors among a large sample of social media users in India (N=5104). We show that peer-to-peer corrections substantially reduce belief in misinformation. Neither motivated reasoning nor the sophistication of these corrective messages conditions their effects. Brief, unsourced and unsubstantiated corrections achieve an effect comparable to that of corrections backed by evidence from credible sources (domain experts and specialized fact-checkers alike). This suggests that merely signaling a doubt about a rumor (regardless of *how* substantiated this signal is) may go a long way in reducing misinformation. These results have implications for both users and platforms.

Keywords: Misinformation; Social Media; Correction; Motivated Reasoning; WhatsApp; India

*This study was registered with Evidence in Governance and Policy (20191008AB) and received IRB approval from Columbia University and IE University (IRB-AAAS3860). The authors thank Ipsa Arora, Ritubhan Gautam, and Hanmant Wanole for research assistance. This research was funded by Facebook. The authors thank Alex Leavitt and Devra Moelher, as well as participants at the Facebook Integrity Research Workshop (June 2019). The manuscript was presented at seminars at Columbia University and Leiden University.

1 Introduction

Social media platforms offer rich grounds for the spread of misinformation (Allcott and Gentzkow 2017). Since the 2016 U.S. election, there has been widespread concern that misinformation on social media distorts public opinion, reduces trust in democracy and even encourages violence. Given its troubling consequences, a vast research agenda has tested and measured the effect of corrections and fact-checking (Flynn, Nyhan, and Reifler 2017; Lazer et al. 2018; Clayton et al. 2019). The majority of this literature focuses on the American context, testing the effect of algorithmic corrections (such as adding a warning label to a disputed message) on belief in misinformation. Such corrections are likely to emanate from a known authority: a media source, a domain expert, or the platform itself. However, these findings may not be representative of many social media contexts in which rumors cannot be corrected by platforms or known authorities. This is especially the case with messaging applications such as WhatsApp, popular in developing countries, where two-way encryption means that algorithmic or authoritative corrections are not possible. Corrective messages must emanate from other users, rather than in a top-down manner, thus the onus to fact-check is on peers themselves.

Consequently, we ask three research questions in this study. First, can misinformation be corrected by peers on messaging applications such as WhatsApp? Second, how substantiated and sophisticated must such peer-to-peer corrections be to have an effect? Third, to what extent are the effects of these corrective messages subject to motivated reasoning? To address these questions, we experimentally evaluate the effect of different types of corrective messages *posted by peers* on belief in misinformation.

We field a survey experiment where respondents are shown a series of hypothetical screenshots of WhatsApp group chats. In each screenshot a user posts a false story to which a peer reacts, with or without a corrective warning, and with or without citing evidence. After seeing each screenshot, respondents are asked to rate the veracity of each rumor. This allows us to evaluate the causal effects of peer-to-peer corrections and of substantiating these corrections. While a handful of studies have focused on peer-to-peer corrections (Bode and Vraga 2018; Vraga and Bode 2018; van der Meer and Jin 2020), we advance this research agenda in a number of important ways. First, we run this study on a more diverse and far larger sample of social media users (N=5104), in the world's largest market for WhatsApp, India. Our study constitutes the first large scale experiment on the prevalence of misinformation among Indian social media users. This setting gives us a chance to evaluate the effect of corrections in a context where levels of education and digital literacy are relatively low, where partisan motivated reasoning may not function as in the American context, and where the misinformation can be a matter of life and death. Second, we issue corrections to a set of diverse false stories on a variety of topics, rather than testing its effect on a single story. Finally, we manipulate the source of the rumor and are thus able to measure whether peer-to-peer corrections (both substantiated or not) can be attenuated by partisan motivated reasoning.

In keeping with existing findings on fact-checking on other social media platforms and

contexts, we confirm that peer-to-peer corrections *are* effective: relative to a no correction condition, respondents exposed to a corrective message from a peer are significantly less likely to believe a given rumor. However, surprisingly, we find that partisan motivated reasoning does not condition the effect of these corrections, suggesting important differences in the mechanisms through which misinformation persists between the American and Indian contexts (Nyhan and Reifler 2010). Finally, contrary to findings from existing studies (Vraga and Bode 2018; van der Meer and Jin 2020), we show that the source and the sophistication of corrective messages do *not* condition their effect: in our experiment, brief, unsourced and unsubstantiated corrections achieve an effect comparable to that of corrections backed by evidence from a variety of credible sources (domain experts and specialized fact-checkers alike).

This suggests that merely signaling a problem with the credibility of a rumor (regardless of *how* sophisticated this signaling is) may go a long way in reducing overall rates of misinformation. Users need not burden themselves with evidence or sourcing in order to affect the beliefs of their peers. While this ability may be socially beneficial when users attempt to correct misperceptions (as in our experiment), this also points to a potential danger: users may lead their peers to be skeptical of true information.

Thus, the implications of our findings are mixed. A naive interpretation may be that users should be encouraged to express their doubts about rumors. Following this idea, a simple solution that tech companies can implement may be to create a "button" for users to express doubt in dubious messages. This would provide a complementary, cost-effective way to limit rates of beliefs in common false stories. However, our results also show that it takes little for users to affect their peers' beliefs in closed group chat settings. As a result, this method may have perverse consequences if true stories are "corrected" for partisan reasons. This suggests that user-driven fact-checking is not a simple "cure" to misinformation on messaging apps, and that platforms will have to develop other, more ambitious strategies, in addition to fact-checking, in order to effectively reduce misinformation.

2 Misinformation On Messaging Applications

For the past few years, technological 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 (Lazer et al. 2018; Barberá et al. 2018), no equivalent solution exists for messaging applications such as WhatsApp, where encryption has been, and is likely to remain, central to the branding of the service.

WhatsApp allows for rapid and private communication within and across groups of users who share pre-existing ties (users need to have each other's phone numbers to communicate). Users rely on Whatsapp in a variety of ways. They use the app to text but also to forward and receive news. WhatsApp allows users to join groups with up to 256 members and share multimedia messages, transforming chat groups into highly active social spaces on which all

types of information circulate. The fact that posts on chat apps are private and protected by encryption means that no-one, including WhatsApp itself, gets to see, read, filter or analyze content, and that it is close to impossible to trace the source or extent of spread of a message in a network.

In many developing countries, political actors have embraced this technology to diffuse misinformation. In India, the site of our study, news reports show that WhatsApp groups formed by the ruling Bharatiya Janta Party (BJP) often morph into havens of misinformation and hateful rhetoric capable of inciting violence (Arun 2019; Farooq 2017; Perrigo 2019). This misinformation is often seen as having electoral consequences, as "the groups are also abuzz with congratulatory messages for Modi and the BJP, even as they cast the opposition parties as anti-Hindu" (Purohit 2019).

Faced with public outcry over violent partisan incidents linked to rumors disseminated through the platform, WhatsApp has turned to an array of solutions. One of these solutions involves encouraging user-driven fact-checking.[1] WhatsApp bought full-page ads in multiple Indian dailies ahead of the 2019 elections (see Figure A.1), asking users to fact-check information themselves to correct fake stories shared on the platform.

To what extent should we expect such a strategy–so far the only known strategy to correct misinformation on encrypted discussion applications–to be effective?

2.1 What Do We Know About Fact-Checking?

Over the past decade, a vast research agenda has explored the effect of providing corrections, warnings, or fact-checking treatments to respondents and consequently measuring their perceived accuracy of news stories. 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, Jackson, and Raj 2017). Chan et al. (2017) find that explicit warnings can reduce the effects of misinformation; Pennycook, Cannon, and Rand (2018) test and find that disputed tags alongside veracity tags can lead to reductions in perceived accuracy; Fridkin, Kenney, and Wintersieck (2015) demonstrate that corrections from professional fact-checkers are more successful at reducing misperceptions.

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 weak, and when outcomes are measured immediately after an intervention, leading to the most significant misperceptions being stable and persistent over time (Flynn, Nyhan, and Reifler 2017). The dominant theoretical explanation for this persistence comes from research on motivated reasoning (Flynn, Nyhan, and Reifler 2017). According to Kunda (1990), when people process information different goals may be activated, including directional goals (trying to reach a desired conclu-

[1] Note that we interchangeably refer to this phenomenon as "user-driven" and as "peer-to-peer" in the rest of this study.

sion) and accuracy goals (trying to process the most correct form of the information). 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 2015). Directional motivated reasoning may thus exacerbate the continued influence of false information even after it has been debunked (Thorson 2016).

2.2 Should Peers Make A Difference?

Studies on misinformation typically rely on experiments in which an authoritative source delivers a corrective message. However, some limited evidence suggests that peer-to-peer corrections can also be effective. While this study is the first to explore such corrections in the context of encrypted messaging applications, a handful of studies have explored the effects of peer corrections as comments and replies to posts on Facebook and Twitter in the American context. Bode and Vraga (2018) compare algorithmic corrections with peer corrections and find that they are both equally effective at dispelling misinformation. van der Meer and Jin (2020) demonstrate that peer corrections are effective at reducing misinformation relative to a control (no correction) condition.

While these studies do little to identify a causal mechanism, a rich literature on source credibility can help explain *why* such peer-to-peer corrections may be effective. Individuals have limited time and cognitive resources to comprehend complex topics such as policy or current affairs, and may therefore use the perceived credibility of sources as a heuristic to guide their evaluation of what is true or false. In general, high-credibility sources are more persuasive and promote greater attitude change than low credibility sources (Eagly and Chaiken 1993). Further, while both expertise and trustworthiness are components of source credibility (Pornpitakpan 2004), the latter is found to be more effective in persuasion than the former (Swire and Ecker 2018). Thus, arguably, peers on average should be seen as more trustworthy than unknown or distant individuals, and users on social media are likely to be able to persuade their peers. Additionally, homophily—the extent to which a person perceives similarities between the way they think and another person does—is often seen as a key determinant of source credibility (Householder and LaMarre 2014). Early investigations of similarity effects (Berscheid 1966; Brock 1965) find that perceived similarity may be an important characteristic of persuasive messaging (Petty, Cacioppo, and Goldman 1981). Since social media networks (and particularly messaging apps such as WhatsApp, for which users need to have each other's phone number to communicate) are built on the basis of homophily /similarity, this may further explain why peer-to-peer corrections would be effective. Accordingly, we hypothesize that peer-to-peer corrections will decrease beliefs in misinformation:

Hypothesis 1: Exposure to corrective information originating from peers will reduce the perceived accuracy of rumors, relative to a no correction condition.

In keeping with findings from the literature on fact-checking, we also hypothesize that

motivated reasoning, specifically partisan motivated reasoning, should attenuate the effects of corrective messages (Taber and Lodge 2006; Nyhan and Reifler 2010). Specifically, we hypothesize that both the political slant of the rumor as well as the news source or politician reporting it can condition the effectiveness of corrections:

Hypothesis 2a: Peer corrections will be more effective when the rumor is attributed to a dissonant politician (compared to when it is unattributed).

Hypothesis 2b: Peer corrections will be less effective when the rumor is attributed to a congenial politician (compared to when it is unattributed).

Hypothesis 3a: Peer corrections will be more effective when the rumor originates from a dissonant media outlet (compared to an unattributed or neutral outlet).

Hypothesis 3b: Peer corrections will be less effective when the rumor originates from a congenial media outlet (compared to an unattributed or neutral outlet).

Hypothesis 4a: Peer corrections will be less effective when the rumor is ideologically congenial to respondents (compared to non-ideological rumors).

Hypothesis 4b: Peer corrections will be more effective when the rumor is ideologically dissonant to respondents (compared to non-ideological rumors).

Existing findings on partisan motivated reasoning mainly come from the American context, where partisan affective polarization notoriously acts as a perceptual screen (Green, Palmquist, and Schickler 2004). In India, partisan affiliations have traditionally been weaker and less stable, so the magnitude of these effects may differ. Nevertheless, since we conduct this experiment during the election when attachments to parties are arguably heightened (Michelitch and Utych 2018), we expect motivated reasoning to play a role in information processing.

2.3 The Role of Substantiation

Corrections emanating from peers could take on a variety of forms. Corrective messages may be short or long; sourced or unsourced; 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 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 all joined in on this corrective effort. Thus users may rely on several different actors to source a corrective message. Empirically, whether citing a source matters for the efficacy of corrections in the Indian context is an open question.

In this study, we distinguish between two types of corrective messages: first, substantiated and sourced corrective messages; second, unsubstantiated and unsourced (and as a result, shorter) messages. There are several reasons to expect that the former (substantiated and sourced messages) have a larger impact. Leading models of persuasive communication emphasize the

potential impact of "perceived argument quality" (Stiff and Mongeau 2016), and find that substantiation increases the perceived quality of an argument, thereby increasing persuasion. In the context of corrections, this suggests that substantiated corrections would be more likely to lead to belief change.

Beyond argument quality, "rational appeals" (i.e. arguments referring to evidence) have also been linked to more persuasive communication. Persuasive speeches contain data, evidence, or supporting materials (German et al. 2016), and scholarship suggests that the most persuasive arguments are ones that provide evidence in the form of information containing three components: a claim, data to support that claim (i.e. evidence), and a warrant that provides a logical connection between the data and the claim (Toulmin 1964). Meta-analyses suggest that "evidence appears to produce general persuasive effects that appear surprisingly stable" (Reinard 1988). In the political science literature, Nyhan and Reifler (2015) identify two distinct types of corrective information: factual elaboration, which places emphasis on facts; and simple, brief rebuttals, using fewer arguments in refuting false information (Lewandowsky et al. 2012). Vraga and Bode (2018) test the effect of social corrections on Facebook and Twitter and find that corrections substantiated with a source are more effective at countering misinformation.

In this paper, we test whether substantiated messages containing an argument and evidence work better than unsubstantiated corrections, where an unknown peer (hereafter referred to as "random guy") shares a rebuttal to false information but provides no justification. Findings from the literature on persuasive communication imply that motivated reasoning and source credibility notwithstanding, persuasive and substantiated messages correcting misinformation that provide data and evidence for the correction should have a greater effect than brief, unsubstantiated messages. Accordingly we hypothesize that:

Hypothesis 5: Exposure to substantiated corrective information will reduce the perceived accuracy of a rumor relative to unsubstantiated corrective information.

3 Design

We test these hypotheses with a survey experiment conducted on a large sample of Hindi speakers (N=5104) living in India.[2]

Our experiment examines the effectiveness of corrective messages on a fictitious but realistic WhatsApp screenshot. We show respondents screenshots of a credible WhatsApp group chat (note that respondents are themselves not a part of the chat, they merely see a fictitious screenshot). In the screen shot, a user posts a false political rumor and a second user (a peer) reacts to the rumor. We include various types of reactions by this second participant and evaluate the effect that different user-posted corrective messages have on belief in common rumors. We

[2]We recruit respondents on Facebook. The ad used to recruit respondents is presented in Figure B.1. Eligibility and exclusion criteria for participants: Hindi speaking Facebook users located in India (age 18+). Due to encryption, it is impossible to recruit directly from WhatsApp. However, 100% of our respondents reported using WhatsApp.

vary the degree of substantiation of the corrective message where peers use a variety of sources used to substantiate the correction. As detailed below, we test for the effect of 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 theoretical priors as to how these different sources might impact beliefs in the Indian context, we do not lay down a formal hypothesis in our pre-analysis plan about the relative impact of these various sources (instead labeling it as a "research question"). Despite this, we compare below whether different sources differently affect belief rates.

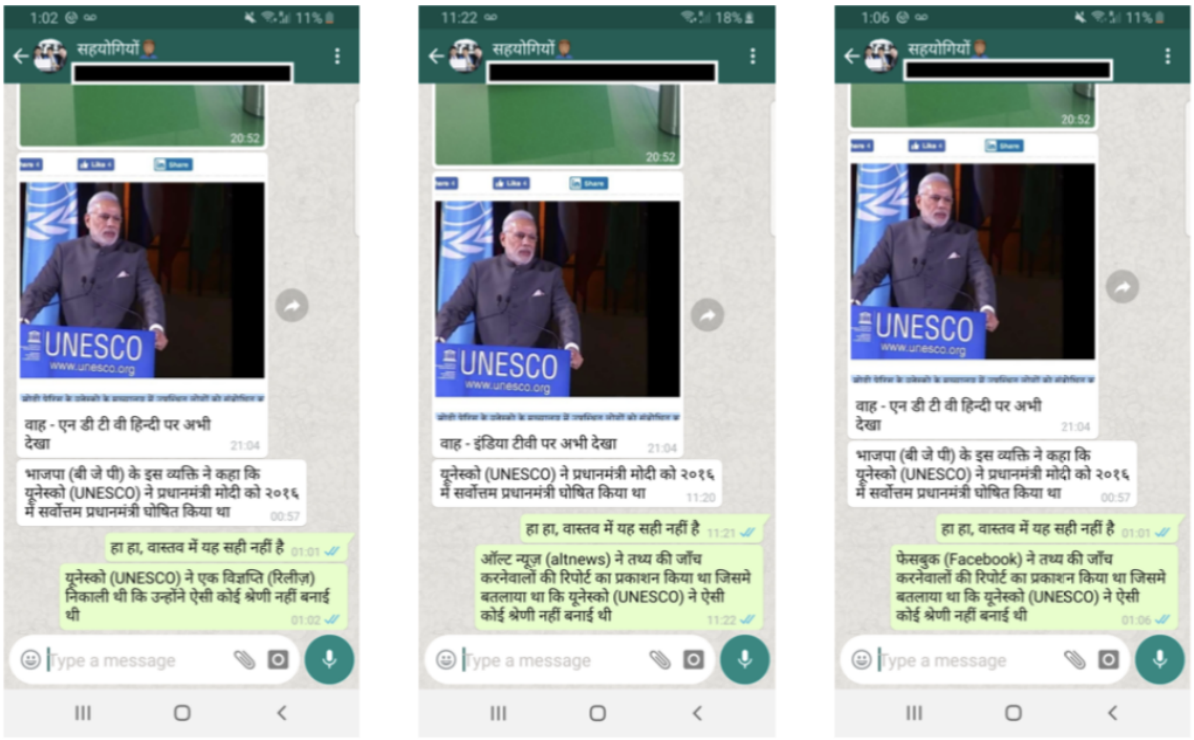

Figure 1: Different versions of a prompt

Figure 1 provides a visual example of the screenshots different respondents were exposed to in the experiment. A user in the chat is shown posting a false story that claims the UNESCO declared the Indian Prime Minister as the "best in the world". In response, a second peer posts a correction to the rumor, varying by experimental group.[3] While we also manipulate various other aspects of the thread (as detailed below), the effect of variations in this last corrective message are the main focus of our empirical analyses.

For each false story, the first participant on the chat posted a visual of a press article (in

[3]Figure C.1 includes the full text of each variation of these corrections in English. The actual experiment was conducted entirely in Hindi.

message 1) and then described the content of the article (in message 2), identifying the publication which reported the story and the politician who made the claim, if applicable. Subsequently, a second participant reacted to these posts, in some cases attempting to correct Participant 1's rumor (through a diverse range of strategies), in other cases simply thanking Participant 1 for the information.

Our study consisted of two sections: a pre-treatment survey section and an experimental section. The pre-treatment section included 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 prompts and outcome measures. Each experimental prompt focused on one rumor that we subsequently asked respondents to evaluate. To ensure respondents became familiar with the format of the prompts and to build up the expectation that some of the stories included in the experiment were true, all respondents began this section by rating the perceived accuracy of a true story (that is, one featuring a verifiable and uncontroversial claim): that Australia holds the record number of victories in the cricket world cup a. The order of the remaining 8 stories was randomized. In total, all respondents evaluated 9 stories, 7 of which were false and 2 of which were true, including the aforementioned Australia story.[5] These stories are listed in Table 1.

Table 1: Dependent Variable Stories

	Story	**Veracity**
1	Australia is the country that has won the ICC cricket world cup the most often	True
2	There is no cure for HIV/AIDS	True
3	In the future, the Muslim population in India will overtake the Hindu population in India	False
4	Polygamy is very common in the Muslim population	False
5	M-R vaccines are associated with autism and retardation	False
6	Drinking cow urine (gomutra) can help build one's immune system	False
7	Netaji Bose did NOT die in a plane crash in 1945	False
8	The BJP has hacked electronic voting machines	False
9	UNESCO declared PM Modi best Prime Minister in 2016	False

These nine stories were chosen following a pretest with an online Indian sample, where we evaluated baseline levels of beliefs in a large list of stories. Each of the stories selected for the final experiment were strongly believed or believed by at least of 25% of the sample of the pretest, with some of these statements being believed by a large *majority* of respondents. Data from this pretest is presented Figures I.1 and I.2.

The final selection of stories was the product of several constraints and choices. To avoid

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

[5] Attrition was less than 1% in our sample.

prompting respondents to systematically reject the veracity of rumors, we included some true stories. But simultaneously, our goal was to maximise respondent exposure to controversial fake political rumors that spread widely during the run up to the 2019 elections in India, hence our distribution skewed in favor of false stories. We selected stories encompassing a broad variety of topics including current electoral politics (stories 8 and 9), health (stories 5 and 6), religion and minorities (stories 3 and 4) and historical conspiracies (story 7).

The WhatsApp discussions respondents were exposed to varied on three dimensions: the media outlet reporting the story and the identity of the politician making the false rumor (in the message of the first participant), and the source and sophistication of the corrective message posted by the second participant.

The first two of these dimensions aim to measure the effect of partisan motivated reasoning. We present false stories emanating from a politician from the Bharatiya Janata Party (BJP), the current ruling party, or from its national-level competitor, the Indian National Congress (INC). Qualitative evidence from the Indian context demonstrates that the BJP dominates the social media environment in India and as a result, political news that goes viral over applications such as WhatsApp often emanates from BJP party sources (Perrigo 2019). To reflect this, we include a majority of stories emanating from BJP politicians, some from INC politicians, as well as from media sources that are traditionally aligned with the two major political parties. We vary the media outlet reporting the story, with some outlets arguably pro-BJP and others anti-BJP or more neutral.

These experimental manipulations are varied in the following manner such that for each story that respondents see on a fictitious screenshot, the story is attributed one of the following media outlets and one of the following partisan conditions.

Experimental variation 1: media outlet reporting false story - 3 possible values:

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

Experimental variation 2: identity of the politician making false stiry- 3 possible values overall:

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

The third dimension aims to measure the effect of the source and sophistication of a correction. Accordingly, we vary the sophistication of the corrective messages that treatment group respondents see. These corrective messages are varied as follows:

Experimental variation 3: Presence/source/sophistication of corrective message on the displayed thread. Respondents were equally likely to be assigned to one of four following correction conditions here, with the fourth condition further subdivided (equally) in five subcategories, as detailed below.

1. No correction (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 (the full text for each experimental manipulation is included in Figure C.1).

2. "Random guy" correction. This consisted of a short and unsourced non-expert correction. As shown in Figure C.1, Participant 2 simply voiced their incredulity with the message "I don't think that's true, bro!" 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 attempted to correct the rumor by citing an authority relevant to the rumor being discussed, for instance a medical professional for health misinformation, or the election commission of India for electoral misinformation.

4. Correction sourced 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 rumor 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 rumor 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 rumor and found it to be erroneous.
    - Facebook – in this case, respondents are told that the online platform Facebook has fact-checked the rumor and found it to be erroneous.
    - WhatsApp – in this case, respondents are told that the online platform WhatsApp has fact-checked the rumor and found it to be erroneous.

Respondents were equally likely to be randomized into one of the four correction conditions listed above, such that about 25% of the sample for each of the 9 rumors received no corrections (but saw screenshots).[6] Respondents who were assigned to read a screenshot (all

[6] A small proportion of respondents (3%) were randomized into a "pure control" condition in which we measure the dependent variable of belief in misinformation without showing respondents any of the screenshots. As demonstrated in Table J.1, we detect no statistical differences in the overall rate of belief in all 7 "false" rumors when comparing our pure control condition (no thread is shown; respondents are only asked to evaluate the veracity of the rumor) to the control condition (WhatsApp thread is shown, but not corrected, after which respondents are asked to evaluate the veracity of the rumor).

except those assigned to the pure control) had an equal probability of being assigned to each of the possible combinations of experimental treatments listed above.

After reading each of the screenshot conversations, respondents answered a single outcome question about their belief in the rumor discussed in the chat:

How accurate is the following statement? [Statement of the rumor]
(very accurate, somewhat accurate, not very accurate, not at all accurate)
Our study took several steps to increase external validity and realism. First, we selected a representative sample of stories, both in terms of thematic focus and diversity of themes. As a point of reference, the false stories presented in (Sinha, Sheikh, and Sidharth 2019) presents a sample of rumors from India relatively comparable to ours. Second, as shown in Figure 1, to avoid biasing responses, we deliberately blacked out the purported names of the participants and presented this as a measure to protect their privacy - hence likely increasing the realism of the experiment. Further, 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). Finally, given that respondents each saw nine screenshots, we varied the specific text of the messages in each screenshot to ensure realism. The spreadsheet presented Figure C.1 gives a full list of treatments for each rumor used in our experiment.

4 Results

Descriptive data from our study demonstrates the high salience of false stories in the Indian context. Despite the fact that the screenshots included a correction in 75% of all conditions, 6 of the 7 false rumors were rated as accurate or somewhat accurate by a *majority* of our sample, as shown in Figure 2.

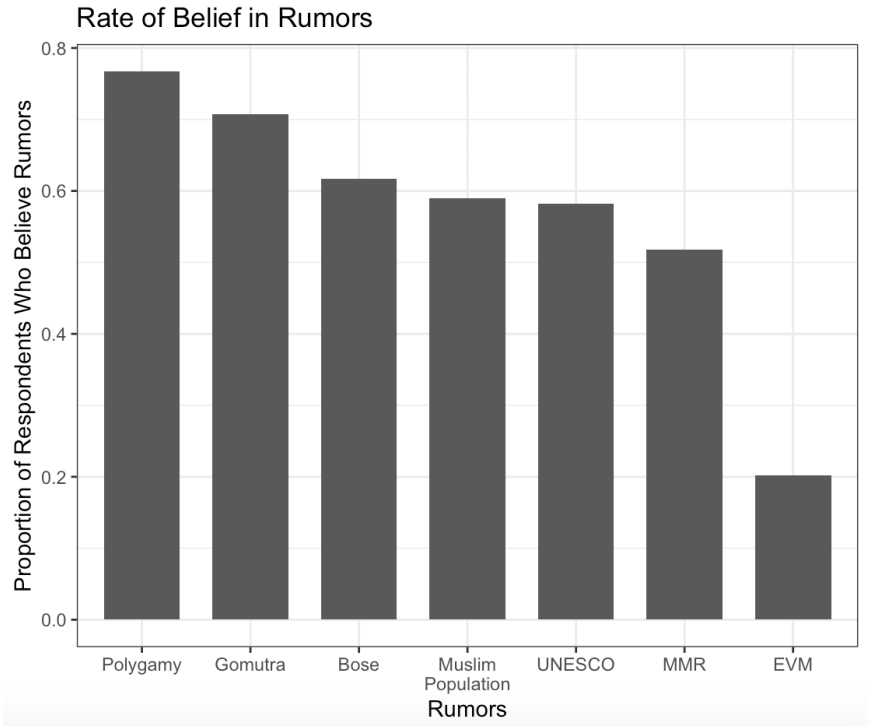

Figure 2: Overall rates of belief in rumors across experimental conditions

4.1 Do Corrective Messages Matter?

Do corrections impact these high levels of belief in rumors? We first present results for Hypothesis 1 which tests whether exposure to corrective messages posted by peers (regardless of their format) reduces the perceived accuracy of rumors.

To test Hypothesis 1, we pool together all the different types of user-driven corrections such that the primary comparison of interest is between having received a correction (of any kind) and not having received a correction. This comparison is expressed in Equation 4.1:

$$Belief\,Accuracy_i = \alpha + \beta_1 AnyCorrection_i + \epsilon_i \tag{4.1}$$

In the equation, i represents the respondent, the *AnyCorrection* variable represents pooled assignment all correction conditions (relative to control). The dependent variable *BeliefAccuracy* measures the self-reported accuracy rating that respondents give to each story on a 4-point scale, with higher values representing more accuracy. We estimate a separate bivariate OLS model for each of the seven false rumors in our experiment, represented by the seven columns in Table 2.[7] In Table 3, we estimate the effect of receiving any correction while controlling for the media

[7]Note that we omit the pure control from these analyses. Combining pure control with the control condition does not change our results, as we demonstrate in Table J.1 that there is no difference between the control and pure control conditions.

source and politician reporting the story. Our model accordingly becomes:

$$BeliefAccuracy_i = \alpha + \beta_1 AnyCorrection_i + \beta_2 DissonantMedia_i + \beta_3 CongenialMedia_i + \beta_4 CopartisanPolitician_i + \beta_5 OutpartisanPolitician_i + \epsilon_i \quad (4.2)$$

Our main result suggests that corrections are effective at reducing beliefs about the accuracy of false news. Exposure to user-driven fact-checking appears to reduce the likelihood that respondents report false rumors to be accurate. Relative to not receiving any correction, receiving a correction (of any form) significantly reduces belief in misinformation, for 6 of 7 false stories (Table 2. This result holds in the presence of added control variables (Table 3. However, we do not obtain a significant result for one false story, the rumor that electronic voting machines (EVMs) were hacked by the BJP ahead of the elections. As Figure 2 demonstrates, belief in this rumor was low to begin with, possibly making it harder for the intervention to have an impact. In contrast, a consistent negative effect appears for the remaining stories, although effect sizes vary across rumors. Particularly, we see effects of larger magnitude (above 0.4 on a scale from 1 to 4) on two of the stories: the MMR vaccine rumor and the UNESCO rumor.

The size of the effect across rumors (Tables 2 and 3) does not seem related to the prior salience of these rumors in our sampled population. As shown in Figures I.1 and I.2, many respondents in our pretest had heard of the widely circulated UNESCO rumor, while fewer had heard of the rumor about MMR vaccines. Yet both led to comparatively large corrective effects. However, this result may manifest as a function of the an artefact of our design. Rumor-specific screenshots varied on dimensions other than the rumor itself: associated image, specific text of the conversation. Since we do not have a way to evaluate the effect of these variations in design across rumor screenshots, we cannot conclude that some topics are more easily corrected than others.

Table 2: Main Effect of Any Correction

	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 3: Main Effect of Any Correction 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

Thus our results suggest that user-driven corrections, the only suitable techniques for encrypted chat applications, can be effective at reducing overall rates of beliefs in patently false rumors circulating on WhatsApp. However, we remain cautious about the scalability of these results. First, owing to the fact that these corrections were issued in a controlled, experimental context, it is possible that the intensity of corrections was stronger in our study than they would be in real life, where WhatsApp groups involve conversations with several people at a time. In spite of this, the finding that corrections by an unidentified peer can consistently produce reductions in misinformed beliefs, despite any lack of incentive to answer correctly, suggests that credible corrections on actual chats built around the principle of homophily would be as or more impactful.

4.2 Motivated Reasoning and Corrections

To what extent are the corrective effects detected above affected by motivated reasoning? Specifically, to what extent are these effects conditional on the party identity of the political actor to whom the rumor is attributed in the first message (Hypotheses 2a and 2b)? To what extent are they conditional on the media outlet on which the rumor is said to have been made (Hypotheses

3a and 3b)? Finally, to what extent are they conditional on political congruence (congeniality) of the rumor itself (Hypotheses 4a and 4b)?

To test Hypotheses 2a and 2b, we limit our analyses to the subset of rumors that are clearly partisan in nature (Rumors 3, 4, 6, 8,and 9) and code whether the rumor 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 self-reported 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 supporters who read an anti-BJP rumor and for INC supporters who read a pro-BJP rumor).

To test Hypotheses 3a and 3b, 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 Hypotheses 4a and 4b, we again limit our analyses to the subset of rumors that are clearly congenial/dissonant to supporters of one of the two major national parties in India: the BJP and INC (rumors 3, 4, 6, 8,and 9). We code rumors 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, Rumors 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 rumors. By contrast, Rumor 8 (EVMs) is coded as dissonant, while Rumors 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, Rumor 8 is coded as congenial while Rumors 3, 4, 6, 9 are coded as dissonant and Rumors 1, 2, 5, and 7 are neither congenial nor dissonant. Note that we again pool members of both major parties in each category (e.g., "dissonant" takes the value of 1 for BJP supporters who read an anti-BJP rumor and for INC supporters who read an anti-INC rumor). Results in Tables G.1 and G.2 underscore this coding choice: we show that rumor congeniality significantly predicts higher rates of belief in rumors. Rumors rated in the pretest as congenial are more likely to be believed, while rumors rated as dissonant are less likely to be believed.

For all these hypotheses, our quantity of interest is the *interaction* between exposure to a corrective message (pooling across all types of corrective messages) and the congeniality of the source/media outlet/rumor. Results from these tests point to a similar conclusion: the effect of

corrections is not limited by partisan motivated reasoning. This is true whether we interact the effect of our corrections with the congeniality of the rumor itself (as per Hypotheses 4a and 4b, tested in Tables 4 and 5), with the partisan leaning of the source of the rumor (as per Hypotheses 2a and 2b, tested Tables D.1 and D.2), or with the the partisan leaning of the media outlet on which the rumor appeared (as per Hypotheses 3a and 3b, tested in tables E.1 and E.2)

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* DissonantClaim	0.062	0.016	-0.060	-0.022	-0.058
	(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

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 do not matter in this context as they frequently do in experiments run on American voters (Nyhan and Reifler 2010), possibly highlighting important differences in the mechanisms through which misinformation persists across countries. Importantly, no such stable interaction exists even among the most clearly ideological and partisan subgroup in our sample: BJP supporters. In Table F.1, we run models in which we interact respondents' level of support for the BJP and the presence of a correction on the thread. In 5 out of 7 cases, we do not detect a 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. These findings underscore the relative absence of partisan motivated reasoning in the Indian context.

4.3 Are Substantiated Corrections More Effective?

While we find that corrections are effective at reducing beliefs in misinformation, we further aim to identify the types of corrections that are most successful. Particularly, we compare substantiated to unsubstantiated messages, and determine whether the source of substantiation plays a role in persuasion.

In Table 6, we rely on OLS models to evaluate the effect of different types of corrections on

the level of belief in each of the rumors, compared to the control condition (omitted category).[8] Figure 3 provides a graphical representation of this regression table. The first row of results in Table 6 represents corrections without substantiation, where a random guy (a peer in the group chat) expresses skepticism with the story but does not explain the source of the skepticism. These corrections take on the form of the message "I don't think that's true, bro!". The remaining rows represent substantiated corrections, but each with a different source of substantiation. While Row 2 contains conditions where the correction was backed by a domain expert, the other rows involve corrections substantiated by various fact-checking actors.

Table 6: 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

Several striking findings emerge from these results. Critically, 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, as seen in Figure 3. This implies that

[8] We similarly exclude the pure control from these analyses. Note, however, that this does not change our results.

the sophistication of user-driven corrections matters very little: 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 rumor is thus as likely to reduce belief in a falsehood as a more carefully crafted correction. Even more striking is the finding that corrective messages substantiated with a domain expert do not 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, as opposed to a domain expert. This further implies that respondents open to belief change do not require much "expertise" in order for their beliefs 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 relative to others. This may suggest that outsourcing fact-checking to credible authorities may not be necessary to improve its overall credibility in this context. Overall, we find that the content of user-driven corrections counts less than the mere existence of a correction.

5 Discussion and Conclusion

Our results confirm that peer-to-peer corrections are effective, as suggested by Bode and Vraga (2018). Our main analyses demonstrate that exposure to a corrective message posted by an unidentified peer on a WhatsApp chat is enough to significantly reduce rates of belief in false information. This is important insofar as respondents in our experiment were not incentivized to pay attention to the message. Besides, they did not know, and by design could not identify, the individual posting this correction. Arguably, such a correction posted on a more homophilic network would achieve a much larger effect.

These results additionally show that corrective effects exist in contexts very different from the United States. Social media users in developing countries - where WhatsApp is the most popular chat application - are different on a number of crucial dimensions. India is a case in point. India has a relatively low literacy and formal education 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. 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 (Allcott and Gentzkow 2017). One may hypothesize that such contexts would make *any* correction more difficult, as readers may not as easily understand the content or the motivation of a detailed correction. We may expect a smaller tendency to change belief after exposure to a correction (in addition to a higher vulnerability on average to misinformation) among populations with lower literacy and lower education. Our results, however, suggest otherwise. Relatively low levels of digital literacy[9] may actually help render corrections effective. Low digital literacy likely implies that

[9]Recent reports demonstrate that the penetration of the Internet in rural India increased from merely 9% in 2015

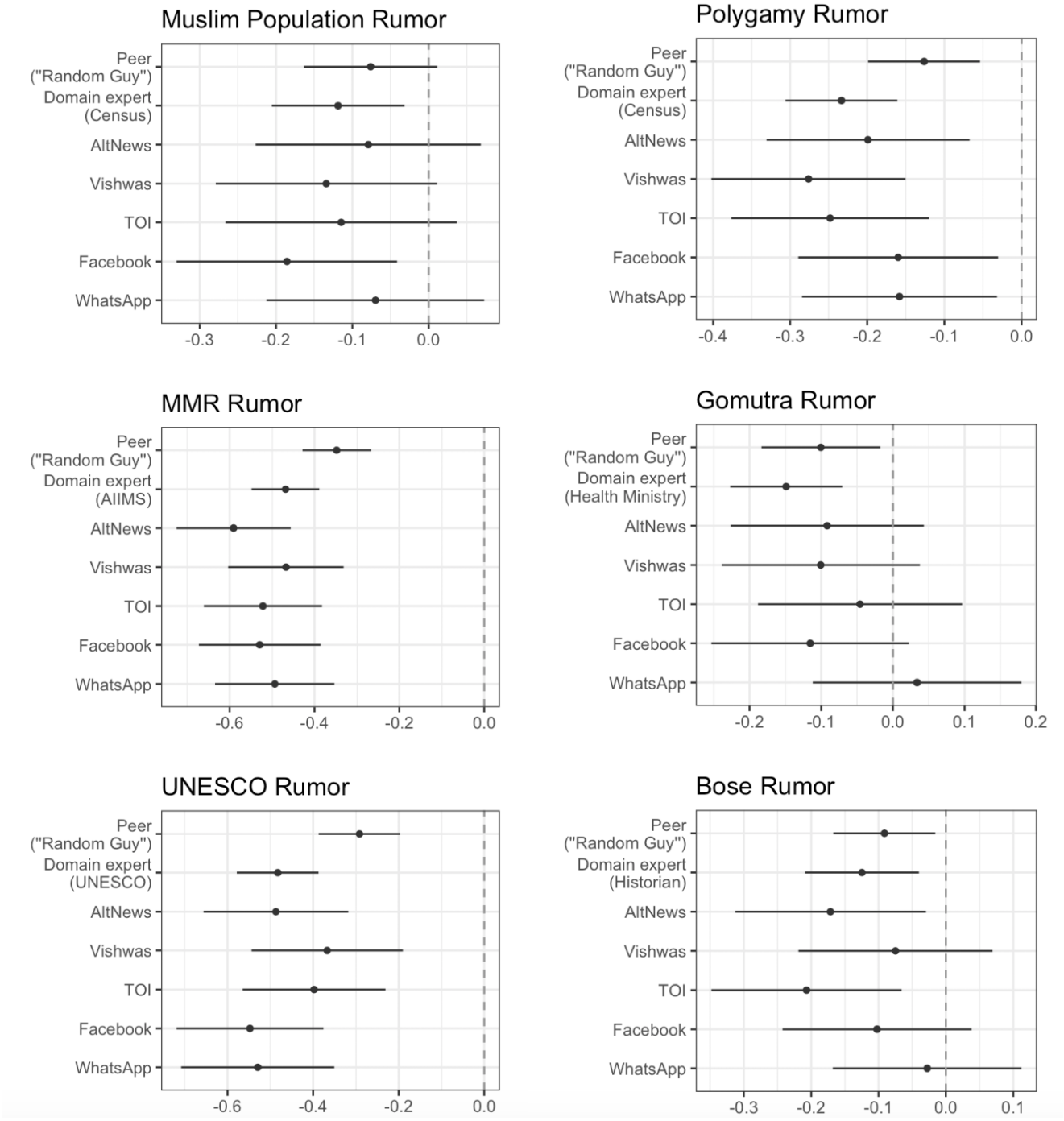

Figure 3: Effect of Correction Source

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, corrections may more easily have a beneficial effect in such contexts.

We do *not*, however, find support for the other hypotheses we tested. In light of findings in the American context, the absence of motivated reasoning in our results is here especially striking. While recent research from India suggests an increasing importance of partisanship as a social identity (Badrinathan 2020; Chhibber and Verma 2018), India is a country that has traditionally had weaker partisan ties, and politics is thought to be more clientelistic rather than programmatic (Ziegfeld 2016). This relative weakness of partisanship may imply that motivated reasoning would *not* constitute as big an obstacle to correcting beliefs, or that motivated reasoning exists in another, non-partisan form (ethnicity or relgion).

The implications of our findings are mixed. These results first confirm that encouraging peer-to-peer fact-checking - as WhatsApp itself has done - could reduce the number of users who believe misinformation. Merely signaling a problem with the credibility of a rumor (regardless of *how* sophisticated this signaling is) may go a long way in reducing rates of beliefs in rumors. This may be seen as good news, as expecting users to post more sophisticated or substantiated corrections may 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.

A possible implication could be that users should be encouraged to effectively "sound off" as easily as possible to express their doubts about rumors posted on the platform. Platforms may do so in a variety of ways. One way to reduce the cost of expressing doubt about a rumor may be to add a simple "button" to express doubt in reference to on-platform rumors, 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 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.

Yet, encouraging users to "sound off" as easily as possible may have perverse consequences. True stories could be "corrected" for partisan reasons, and our results suggest that such misplaced "corrections" by hyper-partisan users are equally likely to be believed. Future research

to 25% in 2018.

should analyse whether a peer posting a factually incorrect correction may have a similar effect on belief change. If so, peer-to-peer corrections may be seen as problematic. More importantly, this would suggest that peer-to-peer fact-checking is not a simple "cure" to misinformation on messaging apps, and that platforms will have to develop other, more ambitious strategies, in addition to fact-checking, in order to effectively reduce misinformation.

Beyond messaging applications, our study opens up broader avenues for research on misinformation in developing countries. Much remains to be uncovered about the ability of misinformation to persuade, and to be corrected, in such settings. The weakness of the *partisan* form of motivated reasoning detected in our study suggests that more comparative work on misperceptions is needed. Future work should explore the psychological mechanisms leading to belief change, and potentially to offline behaviors, especially in countries where the stakes are as high as violence. Such research should also look into information and misinformation processing on encrypted and personal social media networks such as WhatsApp.

References

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

Arun, Chinmayi. 2019. "On WhatsApp, Rumours, Lynchings, and the Indian Government." *Economic & Political Weekly* 54 (6).

Badrinathan, Sumitra. 2020. "Educative Interventions to Combat Misinformation: Evidence From A Field Experiment in India." *Working Paper*. https://sumitrabadrinathan.github.io/Assets/FakeNewsPaper.pdf.

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

Berscheid, Ellen. 1966. "Opinion Change and Communicator-Communicatee Similarity and Dissimilarity." *Journal of Personality and Social Psychology* 4 (6): 670.

Bode, Leticia, and Emily K Vraga. 2018. "See Something, Say Something: Correction of Global Health Misinformation on Social Media." *Health Communication* 33 (9): 1131–1140.

Brock, Timothy C. 1965. "Communicator-Recipient Similarity and Decision Change." *Journal of Personality and Social Psychology* 1 (6): 650.

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

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

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

Eagly, Alice H, and Shelly Chaiken. 1993. *The psychology of attitudes.* Harcourt Brace Jovanovich College Publishers.

Farooq, Gowhar. 2017. "Politics of Fake News: how WhatsApp became a potent propaganda tool in India." *Media Watch* 9 (1): 106–117.

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

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

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

German, Kathleen M, Bruce E Gronbeck, Douglas Ehninger, and Alan H Monroe. 2016. *Principles of public speaking*. New York: Routledge.

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

Housholder, Elizabeth E, and Heather L LaMarre. 2014. "Facebook Politics: Toward a Process Model for Achieving Political Source Credibility Through Social Media." *Journal of Information Technology & Politics* 11 (4): 368–382.

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

Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild et al. 2018. "The science of fake news." *Science* 359 (6380): 1094–1096.

Lewandowsky, Stephan, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest* 13 (3): 106–131.

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

Mosseri, Adam. 2016. "Addressing Hoaxes and Fake News." *Facebook*. December 15, 2016. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

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

Nyhan, Brendan, and Jason Reifler. 2015. "Displacing Misinformation About Events: An Experimental Test of Causal Corrections." *Journal of Experimental Political Science* 2 (1): 81–93.

Pennycook, Gordon, Tyrone D Cannon, and David G Rand. 2018. "Prior Exposure Increases Perceived Accuracy of Fake News." *Journal of Experimental Psychology: General* 147 (12): 1865–1880.

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

Petty, Richard E, John T Cacioppo, and Rachel Goldman. 1981. "Personal Involvement as a Determinant of Argument-Based Persuasion." *Journal of Personality and Social Psychology* 41 (5): 847.

Pornpitakpan, Chanthika. 2004. "The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence." *Journal of Applied Social Psychology* 34 (2): 243–281.

Purohit, Kunal. 2019. "Post CAA, BJP-Linked WhatsApp Groups Mount a Campaign to Foment Communalism." *The Wire*. December 18, 2019. https://thewire.in/media/cab-bjp-whatsapp-groups-muslims/.

Reinard, John C. 1988. "The Empirical Study of the Persuasive Effects of Evidence The Status After Fifty Years of Research." *Human Communication Research* 15 (1): 3–59.

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

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

Stiff, James B, and Paul A Mongeau. 2016. *Persuasive Communication*. New York: Guilford Publications.

Swire, Briony, and Ullrich K H Ecker. 2018. "Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication." *Misinformation and Mass Audiences*: 195–211.

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

Thorson, Emily. 2016. "Belief Echoes: The Persistent Effects of Corrected Misinformation." *Political Communication* 33 (3): 460–480.

Toulmin, Stephen E. 1964. *The Uses of Argument*. Cambrige: Cambridge University Press.

van der Meer, Toni GLA, and Yan Jin. 2020. "Seeking Formula for Misinformation Treatment in Public Health Crises: The Effects of Corrective Information Type and Source." *Health Communication* 35 (5): 560–575.

Vraga, Emily K, and Leticia Bode. 2018. "I do not believe you: how providing a source corrects health misperceptions across social media platforms." *Information, Communication & Society* 21 (10): 1337–1353.

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

Supporting Information for:
"I Don't Think That's True, Bro!"
An Experiment on Fact-checking Misinformation in India

Contents

A 2019 WhatsApp Campaign Promoting User-driven Corrections **2**

B Advertisement Used to Recruit Respondents **3**

C Full Text of Experimental Manipulations **4**

D Hypotheses 2a and 2b **5**

E Hypotheses 3a and 3b **7**

F Heterogeneous Effects of BJP Support **8**

G Main Effect of Congenial / Dissonant Claim **9**

H Summary Statistics **10**

I Pretest Data **14**

J Comparing the Effect of Control vs. Pure Control on Belief in Rumors **18**

A 2019 WhatsApp Campaign Promoting User-driven Corrections

Figure A.1: WhatsApp ad in Indian newspapers

B Advertisement Used to Recruit Respondents

The image below is a screenshot of the advertisement we used to recruit respondents into our survey via Facebook.

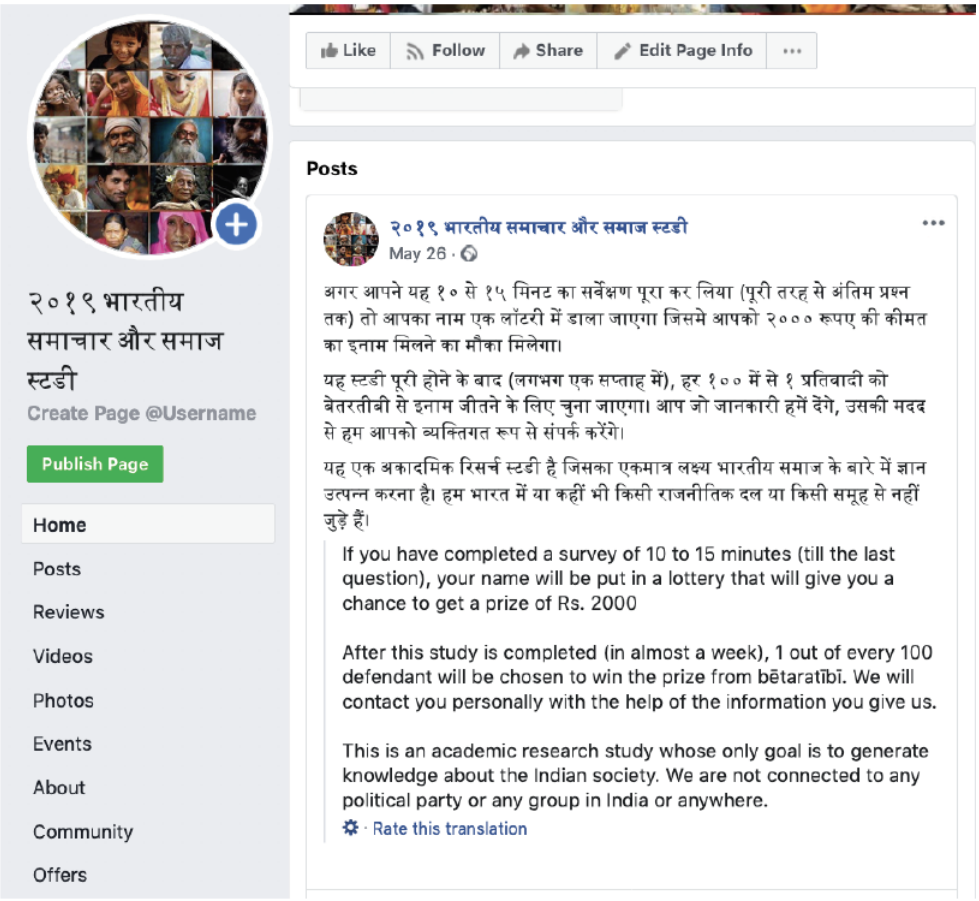

Figure B.1: Sample recruitment ad

C Full Text of Experimental Manipulations

The text displayed on our WhatsApp screenshots to treatment group respondents varied as a function of the type of rumor, the politician / media outlet is is attributed to, and the nature of the correction. The spreadsheet below describes the full text associated with each hypothetical WhatsApp screenshot in each condition.

			SOURCES VARIATIONS				CORRECTION VARIATIONS		
Rumor	Type of claim	Message 1	Message 2 (part 1)	Message 2 (part 2)		Message 3a (control)	Message 3b (Domain	M3b (Factchecker correction)	M3c (random person
1	Congenial to BJP/dissonant to non-BJP	image	Check this out.	In this article published on the NDTV Hindi/Doordarshan/India TV website, a BJP leader claims that the Muslim population in India will overtake the Hindu population in the future.	In this article published on the NDTV Hindi/Doordarshan/India TV website, it is suggested that the Muslim population in India will overtake the Hindu population in India in the future.	thanks -- will read!	The other day I saw a report by prominent demographers that said this was extremely unlikely to happen. So I don't think that's right actually.	The other day I saw a fact-checking report by ALTNEWS/Vishwasnews.com/Times of India/Facebook/WhatsApp that said this was extremely unlikely to happen. So I don't think that's right actually.	The other day I saw something that said this was extremely unlikely to happen. So I don't think that's right actually.
2	Congenial to BJP/dissonant to non-BJP	image	Just saw this!	This appeared on the NDTV Hindi/Doordarshan/India TV website --- A BJP leader explains that polygamy (having several wifes) is very common in the Muslim population.	This appeared on the NDTV Hindi/Doordarshan/India TV website --- it says that polygamy (having several wifes) is very common in the Muslim population.	Thanks. Will check when I have a second.	Hmm, not sure. The Census of India did a report and showed it to be unsubstantiated, though.	Hmm, not sure. A fact-checking report done by ALTNEWS/Vishwasnews.com/Times of India/Facebook/WhatsApp showed this to be unsubstantiated, though.	Hmm, not sure. I saw something showed this to be unsubstantiated, though.
3	Neither congenial nor dissonant to BJP/non-BJP	image	Just came across this article	-	This comes from the NDTV Hindi/Doordarshan/India TV website. Apparently M-R vaccines are associated with autism and retardation.	Wow, ok. will get into this.	Hey I don't think that's true actually. I just saw a report from doctors from AIIMS, there appears to be no basis for this claim...	Hey I don't think that's true actually. I just saw a fact-checking report done by ALTNEWS/Vishwasnews.com/Times of India/Facebook/WhatsApp, there appears to be no basis for this claim...	Hey I don't think that's true actually. Someone told me there was not basis for this claim...
4	Congenial to BJP/dissonant to non-BJP	image	This is worth looking at.	The NDTV Hindi/Doordarshan/India TV website just published this. A bunch of BJP leaders said that drinking cow urine (gomutra) helps build one's immune system.	The NDTV/NDTV Hindi/Doordarshan/RepublicTV/India TV website just published this. Claims that Drinking cow urine (gomutra) helps build one's immune system.	Got it, thanks for sending :)	Actually not sure about this, brother. I saw a report from doctors from AIIMS explaining why this is not correct.	Actually not sure about this, brother. I saw a fact-checking report done by ALTNEWS/Vishwasnews.com/Times of India/Facebook/WhatsApp explaining why this is not correct.	Actually not sure about this, brother. I saw somewhere that this is not correct.
5	Neither congenial nor dissonant to BJP/non-BJP	image	Relevant as the ICC world cup approaches...	-	This comes from the NDTV Hindi/Doordarshan/India TV website. I had forgotten that Australia has more ICC cricket world cup wins than any country!	Great. Thanks for sending :)	-	-	-
6	Neither congenial nor dissonant to BJP/non-BJP	image	Important stuff...	-	the NDTV/NDTV Hindi/Doordarshan/RepublicTV/India TV website published this. sad that there's still no cure for HIV/AIDS	thanks. will definitely read.	-	-	-
7	Congenial to non-BJP/dissonant to BJP	image	Just saw this!	NDTV Hindi/Doordarshan/India TV: several INC leaders claim that the BJP hacks electronic voting machines.	NDTV Hindi/Doordarshan/India TV: some people suggesting that the BJP hacks electronic voting machines.	ok! reading now...	Not sure about this... the Election Commission released a serious report saying there's no basis for this claim	Not sure about this claim. ALTNEWS/Vishwasnews.com/Times of India/Facebook/WhatsApp has come up with a detailed fact-checking report that showed there was no basis for this argument.	Not sure about this claim. I saw somewhere there is no basis for this argument.
8	Congenial to BJP/dissonant to non-BJP	image	Wow	Just saw this on the NDTV Hindi/Doordarshan/India TV website.. This BJP guy said UNESCO declared PM Modi best Prime Minister in 2016.	Just saw this on NDTV Hindi/Doordarshan/India TV website.. UNESCO declared PM Modi best Prime Minister in 2016!	Thanks, boss :)	Haha that's not right actually.. UNESCO put out a release saying they didn't come up with rankings like that.	Haha that's not right actually.. ALTNEWS/Vishwasnews.com/Times of India/Facebook/WhatsApp published a fact-checking thing saying that UNESCO didn't come up with rankings like that.	Haha that's not right actually..
9	Neither congenial nor dissonant to BJP/non-BJP	image	Have a look at this!	From the NDTV Hindi/Doordarshan/India TV website ... Netaji Bose did NOT die in a plane crash in 1945!		wow - thanks for sharing!	This theory has been debunked, I think. I read a report by Delhi University historians explaining there was no ground to believe any of this.	This theory has been debunked, I think. I read a fact-checking report by ALTNEWS/Vishwasnews.com/Times of India/Facebook/WhatsApp explaining there was no ground to believe any of this.	I think this theory has been debunked, though.

Figure C.1: Text for experimental manipulations

D Hypotheses 2a and 2b

Hypothesis 2a: WhatsApp corrections will be more effective when the rumor is attributed to a dissonant politician (compared to an unattributed or neutral politician).

To test this hypothesis, we run the following model:

$$\begin{aligned} BeliefAccuracy_i = \alpha + \beta_1 AnyCorrection_i + \beta_2 DissonantPol_i \\ + \beta_3 AnyCorrection * DissonantPol_i + \epsilon_i \end{aligned} \tag{D.1}$$

As noted in the body of the article, we limit our analyses to the subset of rumors that are clearly partisan in nature (rumors 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. 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).

Table D.1: Effect of Any Correction * Dissonant Speaker on Belief in Rumor

	Dependent variable: Belief in Rumor				
	MuslimPop	Polygamy	Gomutra	EVM	UNESCO
	(1)	(2)	(3)	(4)	(5)
AnyCorrection	-0.127^{***}	-0.178^{***}	-0.092^{***}	-0.007	-0.405^{***}
	(0.038)	(0.031)	(0.033)	(0.033)	(0.040)
DissonantPol	-0.564^{***}	-0.310^{*}	-0.242	-0.229^{**}	-0.184
	(0.149)	(0.164)	(0.154)	(0.108)	(0.203)
AnyCorrection*	0.482^{***}	0.062	-0.055	0.008	0.016
DissonantPol	(0.163)	(0.174)	(0.168)	(0.118)	(0.218)
Constant	2.775^{***}	3.280^{***}	3.018^{***}	1.666^{***}	3.005^{***}
	(0.034)	(0.026)	(0.028)	(0.028)	(0.034)
Observations	5,104	5,103	5,099	5,136	5,109
R^2	0.005	0.012	0.006	0.005	0.022
Adjusted R^2	0.004	0.011	0.006	0.005	0.021
Res. Std. Er.	1.093 (df = 5100)	0.944 (df = 5099)	1.058 (df = 5095)	1.011 (df = 5132)	1.250 (df = 5105)
F Statistic	7.934^{***}	20.647^{***}	10.792^{***}	9.195^{***}	37.700^{***}

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

Hypothesis 2b: WhatsApp corrections will be less effective when the rumor is attributed to a congenial politician (compared to an unattributed or neutral politician).

$$\begin{aligned} Belief\,Accuracy_i = \alpha + \beta_1 AnyCorrection_i + \beta_2 CongenialPol_i + \\ \beta_3 AnyCorrection * CongenialPol_i + \epsilon_i \end{aligned} \tag{D.2}$$

Table D.2: Effect of Any Correction * Congenial Speaker on Belief in Rumor

	Dependent variable: Belief in Rumor				
	MuslimPop	Polygamy	Gomutra	EVM	UNESCO
	(1)	(2)	(3)	(4)	(5)
AnyCorrection	-0.102^{***}	-0.181^{***}	-0.131^{***}	-0.043	-0.400^{***}
	(0.039)	(0.031)	(0.034)	(0.032)	(0.041)
CongenialPol	0.070	0.167	-0.011	0.518^{***}	0.409^{***}
	(0.132)	(0.107)	(0.119)	(0.180)	(0.157)
AnyCorrection*	-0.054	-0.155	0.176	-0.120	-0.349^{**}
CongenialPol	(0.141)	(0.116)	(0.129)	(0.191)	(0.167)
Constant	2.741^{***}	3.262^{***}	3.011^{***}	1.638^{***}	2.979^{***}
	(0.034)	(0.027)	(0.028)	(0.027)	(0.035)
Observations	5,104	5,103	5,099	5,136	5,109
R^2	0.002	0.008	0.004	0.010	0.022
Adjusted R^2	0.001	0.008	0.004	0.009	0.022
Res. Std. Error	1.095 (df = 5100)	0.946 (df = 5099)	1.059 (df = 5095)	1.009 (df = 5132)	1.250 (df = 5105)
F Statistic	2.747^{**}	14.136^{***}	7.207^{***}	16.946^{***}	38.594^{***}
Note:					*p<0.1; **p<0.05; ***p<0.01

E Hypotheses 3a and 3b

Hypothesis 3a: WhatsApp corrections will be more effective when the rumor originates from a dissonant media outlet (compared to an unattributed or neutral outlet).

Hypothesis 3b: WhatsApp corrections will be less effective when the rumor originates from a congenial media outlet (compared to an unattributed or neutral outlet).

To test these hypotheses, 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, New Delhi TV or 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.

We test this hypothesis with the following model:

$$\begin{aligned} BeliefAccuracy_i = \alpha + \beta_1 AnyCorrection_i + \beta_2 CongenialMedia_i + \beta_3 DissonantMedia_i + \\ \beta_4 AnyCorrection * CongenialMedia_i + \beta_5 AnyCorrection * DissonantMedia_i + \epsilon_i \end{aligned} \tag{E.1}$$

Table E.1: Effect of Any Correction * Media Outlet Source on Belief in Rumor

	Dependent variable: Belief in Rumor						
	MuslimPop	Polygamy	MMR	Gomutra	EVM	UNESCO	Bose
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
AnyCorrection	−0.150***	−0.167***	−0.419***	−0.128***	−0.006	−0.399***	−0.110***
	(0.041)	(0.033)	(0.036)	(0.036)	(0.035)	(0.043)	(0.034)
Congenial	−0.224*	0.230**	−0.078	0.080	−0.163	0.036	0.186
Media	(0.126)	(0.113)	(0.121)	(0.108)	(0.111)	(0.151)	(0.114)
Dissonant	−0.132	0.061	−0.050	−0.184	0.051	0.094	−0.048
Media	(0.121)	(0.108)	(0.121)	(0.116)	(0.118)	(0.143)	(0.116)
AnyCorrection*	0.271**	−0.194	−0.049	−0.056	0.080	−0.067	−0.174
CongenialMedia	(0.135)	(0.122)	(0.131)	(0.119)	(0.121)	(0.162)	(0.125)
AnyCorrection*	0.219*	−0.134	−0.067	0.272**	−0.134	−0.092	0.091
DissonantMedia	(0.131)	(0.116)	(0.130)	(0.126)	(0.127)	(0.155)	(0.127)
Constant	2.773***	3.256***	2.834***	3.015***	1.658***	2.992***	2.781***
	(0.036)	(0.027)	(0.030)	(0.029)	(0.029)	(0.036)	(0.027)
Observations	5,104	5,103	5,061	5,099	5,136	5,109	5,117
R^2	0.003	0.009	0.038	0.003	0.002	0.021	0.003
Adjusted R^2	0.002	0.008	0.037	0.002	0.001	0.020	0.002
Res. Std. Er.	1.095	0.945	1.038	1.060	1.013	1.251	1.048
F Statistic	3.055***	9.650***	39.456***	3.394***	1.663	21.697***	3.190***

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

F Heterogeneous Effects of BJP Support

To complement our tests of motivated reasoning (based on the congeniality/dissonance of the information presented and the source of the information), we present OLS results from models that test whether BJP voters react differently to corrective information.

$$\begin{aligned} Belief\,Accuracy_i = \alpha + \beta_1 AnyCorrection_i + \beta_2 BJPSupport_i \\ + \beta_3 AnyCorrection * BJPSupport_i + \epsilon_i \end{aligned} \tag{F.1}$$

Table F.1: Effect of BJP Support * Correction

	Dependent variable: Belief in Rumor						
	MuslimPop	Polygamy	MMR	Gomutra	EVM	UNESCO	Bose
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
AnyCorrection	−0.078	−0.168***	−0.452***	−0.027	−0.123**	−0.405***	−0.058
	(0.064)	(0.053)	(0.058)	(0.056)	(0.052)	(0.068)	(0.054)
BJP Support	0.424***	0.338***	0.092	0.578***	−0.870***	0.491***	0.237***
	(0.069)	(0.055)	(0.060)	(0.057)	(0.054)	(0.071)	(0.053)
AnyCorrection *	−0.043	−0.027	0.006	−0.114*	0.151**	−0.017	−0.081
BJP Support	(0.078)	(0.064)	(0.070)	(0.068)	(0.064)	(0.083)	(0.066)
Constant	2.460***	3.040***	2.765***	2.616***	2.238***	2.669***	2.629***
	(0.057)	(0.045)	(0.050)	(0.047)	(0.045)	(0.058)	(0.044)
Observations	5,104	5,103	5,061	5,099	5,136	5,109	5,117
R^2	0.029	0.032	0.037	0.051	0.124	0.052	0.009
Adjusted R^2	0.029	0.032	0.037	0.050	0.123	0.051	0.009
Res. Std. Er.	1.080	0.934	1.038	1.034	0.949	1.231	1.045
F Statistic	51.459***	56.702***	65.187***	90.418***	241.189***	93.117***	16.129***
Note:							*p<0.1; **p<0.05; ***p<0.01

G Main Effect of Congenial / Dissonant Claim

In this section, we show that the claims we code as congenial to respondents are more likely to be believed (G.1) and that the claims we code as dissonant to respondents are less likely to be believed (G.2). In each case we run a simple bivariate OLS model:

$$Belief = \alpha + \beta_1(CongenialClaim/DissonantClaim) + \epsilon \tag{G.1}$$

Table G.1: Effect of Rumor Congeniality on Belief

	Dependent variable: Belief in Rumor				
	MuslimPop	Polygamy	Gomutra	EVM	UNESCO
	(1)	(2)	(3)	(4)	(5)
CongenialClaim	0.218***	0.214***	0.378***	0.480***	0.309***
	(0.031)	(0.027)	(0.030)	(0.029)	(0.036)
Constant	2.530***	3.001***	2.702***	1.478***	2.506***
	(0.025)	(0.021)	(0.024)	(0.017)	(0.028)
Observations	5,104	5,103	5,099	5,136	5,109
R^2	0.009	0.012	0.030	0.049	0.014
Adjusted R^2	0.009	0.012	0.030	0.049	0.014
Res. Std. Er.	1.091 (df = 5102)	0.944 (df = 5101)	1.045 (df = 5097)	0.989 (df = 5134)	1.255 (df = 5107)
F Statistic	48.141***	62.228***	157.494***	264.374***	73.238***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table G.2: Effect of Rumor Dissonance on Belief

	Dependent variable: Belief in Rumor				
	MuslimPop	Polygamy	Gomutra	EVM	UNESCO
	(1)	(2)	(3)	(4)	(5)
DissonantClaim	−0.157***	−0.176***	−0.309***	−0.625***	−0.231***
	(0.033)	(0.028)	(0.031)	(0.028)	(0.038)
Constant	2.716***	3.190***	3.035***	2.019***	2.772***
	(0.019)	(0.016)	(0.018)	(0.022)	(0.021)
Observations	5,104	5,103	5,099	5,136	5,109
R^2	0.004	0.008	0.019	0.090	0.007
Adjusted R^2	0.004	0.007	0.018	0.089	0.007
Res. Std. Er.	1.093 (df = 5102)	0.946 (df = 5101)	1.051 (df = 5097)	0.967 (df = 5134)	1.259 (df = 5107)
F Statistic	22.998***	38.607***	96.136***	505.256***	37.676***
Note:					*p<0.1; **p<0.05; ***p<0.01

H Summary Statistics

Table H.1: Summary Statistics for Muslim Population Rumor

Statistic	N	Mean	St. Dev.	Min	Median	Max
Belief in Rumor	5,104	2.665	1.096	1	3	4
Any Correction	5,104	0.781	0.414	0	1	1
Outpartisan Speaker	5,104	0.069	0.253	0	0	1
Copartisan Speaker	5,104	0.126	0.332	0	0	1
Congenial Media	5,104	0.134	0.341	0	0	1
Dissonant Media	5,104	0.127	0.333	0	0	1
Congenial Claim	5,104	0.616	0.486	0	1	1
Dissonant Claim	5,104	0.326	0.469	0	0	1
BJP Partisan	5,104	0.678	0.467	0	1	1
Congress Partisan	5,104	0.057	0.233	0	0	1
Pure Control	5,104	0.023	0.148	0	0	1
Peer Correction	5,104	0.258	0.438	0	0	1
Expert Correction	5,104	0.261	0.439	0	0	1
Alt News	5,104	0.051	0.220	0	0	1
Vishwas	5,104	0.053	0.225	0	0	1
TOI	5,104	0.048	0.213	0	0	1
Facebook	5,104	0.054	0.226	0	0	1
WhatsApp	5,104	0.056	0.229	0	0	1

Table H.2: Summary Statistics for Polygamy Rumor

Statistic	N	Mean	St. Dev.	Min	Median	Max
Belief in Rumor	5,103	3.133	0.949	1	3	4
Any Correction	5,103	0.735	0.441	0	1	1
Outpartisan Speaker	5,103	0.063	0.243	0	0	1
Copartisan Speaker	5,103	0.118	0.323	0	0	1
Congenial Media	5,103	0.116	0.321	0	0	1
Dissonant Media	5,103	0.127	0.333	0	0	1
Congenial Claim	5,103	0.618	0.486	0	1	1
Dissonant Claim	5,103	0.323	0.468	0	0	1
BJP Partisan	5,103	0.680	0.467	0	1	1
Congress Partisan	5,103	0.058	0.234	0	0	1
Pure Control	5,103	0.022	0.145	0	0	1
Peer Correction	5,103	0.247	0.431	0	0	1
Expert Correction	5,103	0.247	0.431	0	0	1
AltNews	5,103	0.045	0.208	0	0	1
Vishwas	5,103	0.050	0.219	0	0	1
TOI	5,103	0.048	0.215	0	0	1
Facebook	5,103	0.047	0.212	0	0	1
WhatsApp	5,103	0.050	0.218	0	0	1

Table H.3: Summary Statistics for MMR Rumor

Statistic	N	Mean	St. Dev.	Min	Median	Max
Belief in Rumor	5,061	2.500	1.058	1	3	4
Any Correction	5,061	0.729	0.444	0	1	1
Congenial Media	5,061	0.125	0.331	0	0	1
Dissonant Media	5,061	0.121	0.326	0	0	1
BJP Partisan	5,061	0.680	0.466	0	1	1
Congress Partisan	5,061	0.058	0.234	0	0	1
Pure Control	5,061	0.022	0.146	0	0	1
Peer Correction	5,061	0.234	0.424	0	0	1
Expert Correction	5,061	0.243	0.429	0	0	1
AltNews	5,061	0.054	0.227	0	0	1
Vishwas	5,061	0.053	0.224	0	0	1
TOI	5,061	0.050	0.218	0	0	1
Facebook	5,061	0.046	0.210	0	0	1
WhatsApp	5,061	0.049	0.215	0	0	1

Table H.4: Summary Statistics for Gomutra Rumor

Statistic	N	Mean	St. Dev.	Min	Median	Max
Belief in Rumor	5,099	2.935	1.061	1	3	4
Any Correction	5,099	0.706	0.455	0	1	1
Outpartisan Speaker	5,099	0.061	0.240	0	0	1
Copartisan Speaker	5,099	0.121	0.326	0	0	1
Congenial Media	5,099	0.128	0.334	0	0	1
Dissonant Media	5,099	0.127	0.334	0	0	1
Congenial Claim	5,099	0.618	0.486	0	1	1
Dissonant Claim	5,099	0.323	0.468	0	0	1
BJP Partisan	5,099	0.680	0.467	0	1	1
Congress Partisan	5,099	0.058	0.233	0	0	1
Pure Control	5,099	0.023	0.151	0	0	1
Peer Correction	5,099	0.204	0.403	0	0	1
Expert Correction	5,099	0.251	0.433	0	0	1
AltNews	5,099	0.053	0.224	0	0	1
Vishwas	5,099	0.051	0.221	0	0	1
TOI	5,099	0.048	0.214	0	0	1
Facebook	5,099	0.052	0.222	0	0	1
WhatsApp	5,099	0.046	0.209	0	0	1

Table H.5: Summary Statistics for EVM Rumor

Statistic	N	Mean	St. Dev.	Min	Median	Max
Belief in Rumor	5,136	1.633	1.014	1	1	4
Any Correction	5,136	0.728	0.445	0	1	1
Outpartisan Speaker	5,136	0.125	0.331	0	0	1
Copartisan Speaker	5,136	0.063	0.243	0	0	1
Congenial Media	5,136	0.125	0.331	0	0	1
Dissonant Media	5,136	0.125	0.331	0	0	1
Congenial Claim	5,136	0.323	0.468	0	0	1
Dissonant Claim	5,136	0.618	0.486	0	1	1
Congress Partisan	5,136	0.057	0.232	0	0	1
BJP Partisan	5,136	0.680	0.467	0	1	1
Pure Control	5,136	0.022	0.148	0	0	1
Peer Correction	5,136	0.250	0.433	0	0	1
Expert Correction	5,136	0.221	0.415	0	0	1
AltNews	5,136	0.051	0.220	0	0	1
Vishwas	5,136	0.053	0.224	0	0	1
TOI	5,136	0.051	0.220	0	0	1
Facebook	5,136	0.055	0.227	0	0	1
WhatsApp	5,136	0.048	0.214	0	0	1

Table H.6: Summary Statistics for UNESCO Rumor

Statistic	N	Mean	St. Dev.	Min	Median	Max
Belief in Rumor	5,109	2.697	1.264	1	3	4
Any Correction	5,109	0.734	0.442	0	1	1
Outpartisan Speaker	5,109	0.059	0.235	0	0	1
Copartisan Speaker	5,109	0.118	0.322	0	0	1
Congenial Media	5,109	0.119	0.324	0	0	1
Dissonant Media	5,109	0.119	0.323	0	0	1
Congenial Claim	5,109	0.619	0.486	0	1	1
Dissonant Claim	5,109	0.324	0.468	0	0	1
Congress Partisan	5,109	0.057	0.233	0	0	1
BJP Partisan	5,109	0.681	0.466	0	1	1
Pure Control	5,109	0.010	0.099	0	0	1
Peer Correction	5,109	0.253	0.435	0	0	1
Expert Correction	5,109	0.249	0.433	0	0	1
AltNews	5,109	0.049	0.215	0	0	1
Vishwas	5,109	0.044	0.204	0	0	1
TOI	5,109	0.050	0.218	0	0	1
Facebook	5,109	0.047	0.211	0	0	1
WhatsApp	5,109	0.042	0.202	0	0	1

Table H.7: Summary Statistics for Bose Rumor

Statistic	N	Mean	St. Dev.	Min	Median	Max
Belief in Rumor	5,117	2.716	1.049	1	3	4
Any Correction	5,117	0.663	0.473	0	1	1
Congenial Media	5,117	0.126	0.331	0	0	1
Dissonant Media	5,117	0.114	0.317	0	0	1
BJP Partisan	5,117	0.681	0.466	0	1	1
Congress Partisan	5,117	0.057	0.232	0	0	1
Pure Control	5,117	0.023	0.149	0	0	1
Peer Correction	5,117	0.252	0.434	0	0	1
Expert Correction	5,117	0.175	0.380	0	0	1
AltNews	5,117	0.047	0.211	0	0	1
Vishwas	5,117	0.045	0.207	0	0	1
TOI	5,117	0.047	0.212	0	0	1
Facebook	5,117	0.048	0.214	0	0	1
WhatsApp	5,117	0.048	0.214	0	0	1

I Pretest Data

We ran a pretest on a panel of Facebook-recruited Indian respondents in early May 2019 (N=640) to measure the salience and rate of belief in 37 different rumors commonly disseminated on social media in India. These rumors were:

1. In the future, the Muslim population in India will overtake the Hindu population in India.
2. Polygamy is very common in the Muslim population.
3. Papaya leaf juice is a good way to cure dengue fever.
4. The food prepared by menstruating women is contaminated and rots faster.
5. M-R vaccines are associated with autism and retardation.
6. M-R vaccines are sometimes used by the government to control the population growth amongst certain groups.
7. One must sleep on the left side after having food, as any other sleeping position could be harmful to the digestive tract.
8. Drinking cow urine (gomutra) can help build one's immune system.
9. Gandhi did not try to save Baghat Singh and may even have been a co-conspirator in his death.
10. Indira Gandhi converted to Islam after marrying Feroze Gandhi.
11. Netaji Bose did NOT die in a plane crash in 1945.
12. Arvind Kejriwal has a drinking problem and makes videos while drunk.
13. Sonia Gandhi smuggled Indian treasures to Italy.
14. The BJP has hacked electronic voting machines.
15. NRIs will be able to vote online during the 2019 elections.
16. New Indian notes have a GPS chip to detect black money.
17. UNESCO declared PM Modi best Prime Minister in 2016.
18. WhatsApp profile pictures can be used by ISIS for terror activities.
19. People with cancer shouldn't eat sugar as it feeds cancer cells.
20. Biopsy causes a tumour to turn cancerous.
21. One should not take the P/500 paracetamol, as doctors have shown it to contain machupo, one of the most dangerous viruses in the world.
22. Dengue can be prevented with coconut oil, cardamom seeds, and eupatorium perfoliatum.
23. Amul Kulfi has some pig contents.
24. Drinking Pepsi after eating Polo or Mentos can cause instant death.
25. The BJP is in league with Facebook to remove anti-BJP pages and advertisements.
26. PM Modi hired a makeup artist for 15 lakh monthly salary.
27. Amit Shah personally ordered the assassination of Judge Loya.
28. Arun Jaitley is the current minister of Finance of the Government of India.
29. Scientists warn that current air quality in Delhi shortens lifespan by several years on average.
30. Pryanka Chopra married an American singer in 2018.
31. Mukesh Ambani's residence in Mumbai is the largest private home in the world.

32. India is now the fifth largest economy in the world.
33. Sachin Tendulkar owns the record number of runs record in the ICC cricket world cup.
34. Australia is the country that has won the ICC cricket world cup the most often.
35. According to the 2011 census, Sikhs represent less than 2% of the total Indian population.
36. There is no vaccine that cures HIV/AIDS.
37. Gandhi started his political career in South Africa before coming back to India.

In Figure I.1 we plot we percent of the pretest sample who said they heard each rumor. In Figure I.2 we plot the percent of the sample who said a given rumor was very accurate or somewhat accurate. We highlight the rumors from this list that we selected for the final experiment.

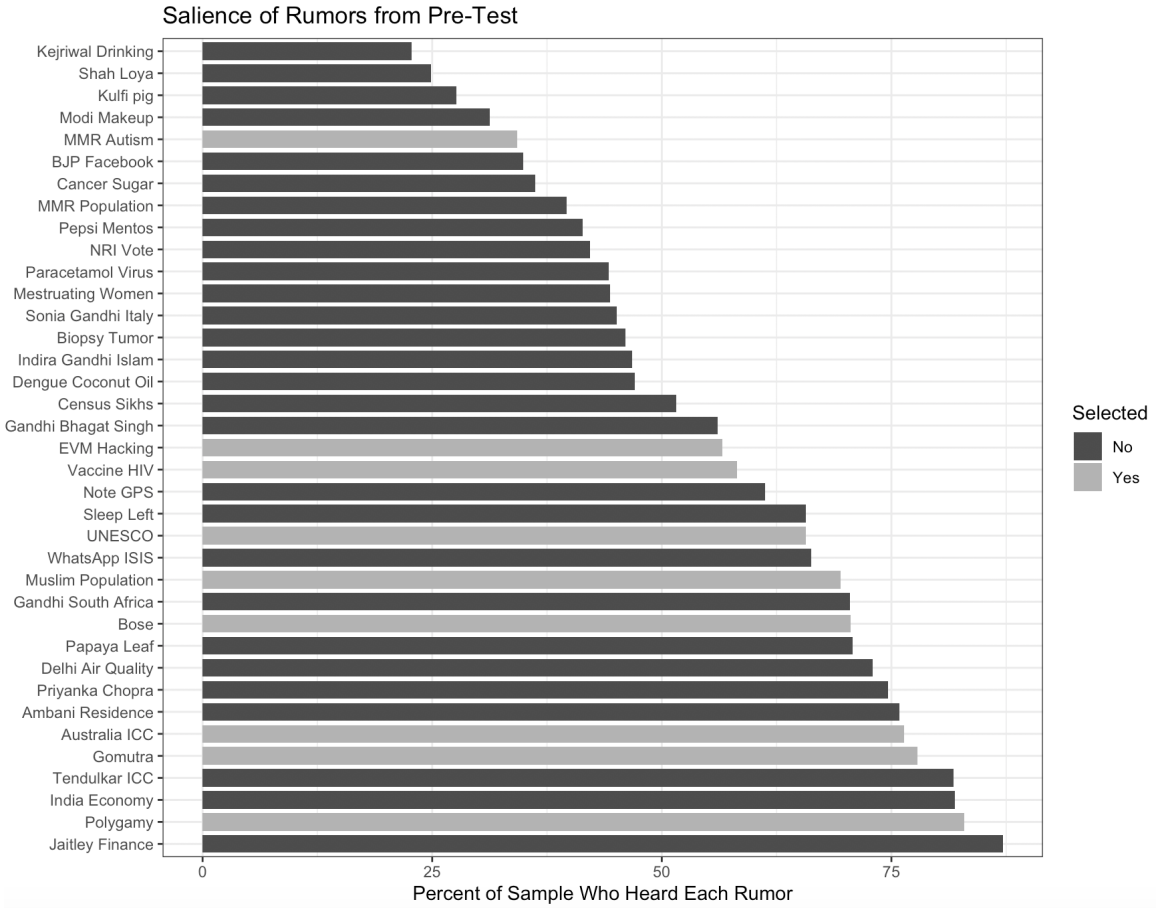

Figure I.1: Salience of Pretest Rumors

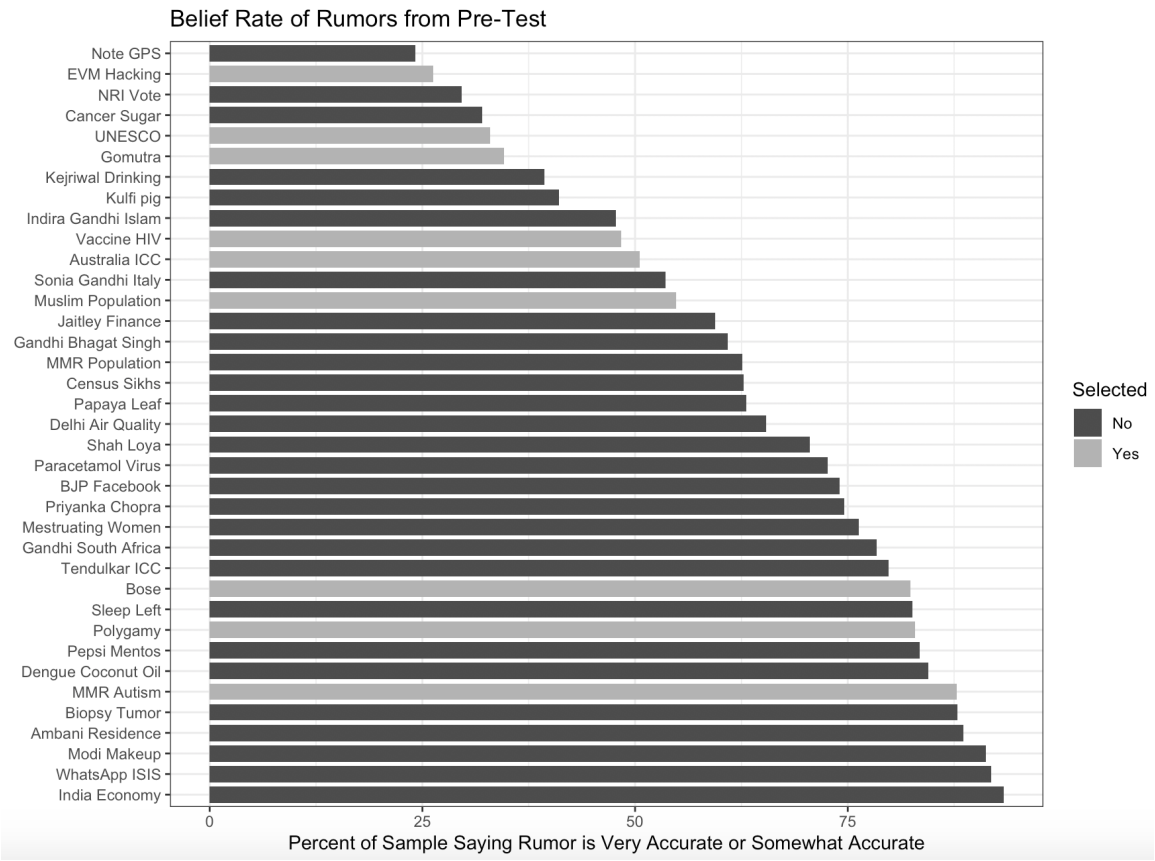

Figure I.2: Belief in Pretest Rumors

J Comparing the Effect of Control vs. Pure Control on Belief in Rumors

In this section, we restrict our sample to items for which respondents received either the control condition ("neutral" reaction by a second user but no correction) or the pure control (no screenshot; respondents directly asked the dependent variable). In the regressions presented below, we test in this sub-sample the effect of receiving the "pure control", compared to the control condition, which is here the omitted category. We run a simple bivariate OLS model where the independent variable is an indicator representing assignment to pure control. We find no differences between the control and pure control conditions.

Table J.1: Difference Between Control and Pure Control Conditions

	Dependent variable: Belief in Rumor						
	MuslimPop	Polygamy	MMR	Gomutra	EVM	UNESCO	Bose
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Pure Control	−0.172	−0.059	−0.028	0.100	−0.074	0.085	−0.084
	(0.108)	(0.089)	(0.100)	(0.102)	(0.100)	(0.172)	(0.102)
Constant	2.763***	3.277***	2.829***	3.001***	1.656***	2.993***	2.834***
	(0.035)	(0.025)	(0.028)	(0.029)	(0.029)	(0.035)	(0.030)
Observations	1,117	1,351	1,371	1,458	1,398	1,260	1,382
R^2	0.002	0.0003	0.0001	0.001	0.0004	0.0002	0.0005
Adjusted R^2	0.001	−0.0004	−0.001	−0.00002	−0.0003	−0.001	−0.0002
Res. Std. Er.	1.097	0.898	1.008	1.061	1.032	1.205	1.055
F Statistic	2.539	0.437	0.076	0.972	0.538	0.244	0.675

Note: *p<0.05; **p<0.01; ***p<0.001