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

An Experiment on Fact-checking Misinformation in India*

Sumitra Badrinathan
University of Oxford

Simon Chauchard
Leiden University

Abstract

In many developing countries, misinformation spreads on private online spaces such as WhatsApp – built on norms of solidarity, and often appropriated by political actors. But since private platforms like WhatsApp cannot control content, can the users of these platforms intervene to curb the spread of misinformation? This study uses a large-scale experiment to evaluate the effect of different types of social corrections on the persistence of misinformation in India (N=5100). We show that social corrections substantially reduce beliefs in misinformation. Importantly, these positive effects are not attenuated by partisan motivated reasoning, highlighting a striking difference from Western contexts. However, we find that the presence of a correction matters more than how sophisticated it is: brief, unsourced and unsubstantiated corrections achieve an effect comparable to that of corrections backed by evidence from credible sources. These results have implications for both users and platforms.[1]

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

*Manuscript currently under review.

[1]Registration for this study is available at EGAP.

In September 2018, the president of the Bharatiya Janta Party (BJP)– the ruling, right-leaning party in India– Amit Shah, addressed the party's social media volunteers who ran WhatsApp groups for their constituents. "We are capable of delivering any message we want to the public," he said, "whether sweet or sour, truth or lie." He was not overselling the party's strategy. Ahead of national elections in April and May 2019, India's political parties, and particularly the BJP, poured money into creating hundreds of thousands of WhatsApp group chats to spread political messages, often times comprising of misinformation (Perrigo 2019).

The success of this political strategy relied on the growing appeal of private online group chat spaces such as WhatsApp. During the 2019 elections in India, mass forwarded WhatsApp messages were a key conduit through which political misinformation, along with waves of falsehoods and hysteria, spread to vast swaths of a population relatively new to social media. Political parties capitalized on the nature of the platform's psychological appeal (Garimella and Eckles 2020)– while the app's privacy meant that nobody outside of a group's members would be privy to its conversations, the intimacy it offered created a haven for misinformation. WhatsApp groups are created for seemingly innocuous motives – parent teacher associations, neighborhood residents groups, mutual aid – but they can promote a sense of solidarity, making the misinformation on them more likely to be trusted (Davies 2020). While such misinformation can often be insidious, in India it has been linked to polarization, mobilization to violence, and even murder (Ara 2020).

The rise in popularity of WhatsApp around the developing world, and its appropriation by nefarious political actors, poses an important empirical question. What can be done to counter misinformation on private online spaces where people are organized into trusted communities built on norms of solidarity?

To answer this question, we implement a large-scale online experiment with over 5,100 respondents in India, the largest market for WhatsApp in the world, in the after-

math of the 2019 elections. While a vast research agenda has tested and measured the effect of corrections and fact-checking, nearly all of this extent literature focuses on the U.S. and other developed democracies where misinformation spreads via public sites such as Facebook and Twitter. Solutions in these contexts are not easily adapted for misinformation distributed on encrypted chat applications such as WhatsApp, where no one, including the app developers themselves, can see, read or analyze messages. The private nature of WhatsApp implies that the onus of debunking and correcting misperceptions falls on users themselves: they must decide to contradict other users, and hope that their intervention influences the subsequent information processing.

To this end, we ask whether social corrections, or peers correcting each other on WhatsApp groups, can reduce the uptake of misinformation. We implement an experiment where respondents are shown a series of hypothetical WhatsApp group chat conversations. In each conversation, a user posts a false story to which a peer reacts, with or without a correction, and with or without citing evidence. Our experiment demonstrates that the presence of a social correction can significantly improve information processing. Relative to a no correction condition, witnessing any type of social correction reduces the perceived accuracy of beliefs in misinformation. Importantly, this effect is robust to respondents' party identity. Contrary to much of the scholarship on corrections (Nyhan and Reifler 2010), we find that partisan motivated reasoning does not attenuate these corrective effects, suggesting important differences in the mechanisms through which misinformation persists in contexts like India. Finally, we show that the source and the sophistication of corrective messages do not condition their effect: in our experiment, brief, unsourced and unsubstantiated corrective messages achieve an effect comparable to that of corrections citing evidence from a variety of credible sources.

To our knowledge, this is the first experiment to test the effect of any sort of correction strategy to temper the uptake of misinformation in India, and the first to systematically test the effect of corrective strategies on WhatsApp. While our findings confirm

that corrective messages work, they challenge the claim that partisan motivated reasoning necessarily affects information processing. This study hopes to spark a research agenda tailored to the study of misinformation on private spaces such as WhatsApp. We demonstrate that not only are the solutions required for such contexts different, but also that the mechanisms of belief in misinformation may depart from theoretical expectations.

Correcting Misinformation: Beyond the U.S. Context

Over the past decade, a vast research agenda has tested solutions to fight misinformation on social media. These include corrective interventions with warnings or fact-checking treatments. 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 successful at reducing misperceptions. Solutions to fight misinformation also rely on inoculating respondents against false news. Such treatments rely on reminding respondents to be critical consumers of news (Tully, Vraga, and Bode 2020; Vraga, Bode, and Tully 2020), presenting mainstream scientific views alongside contrarian views (Cook, Lewandowsky, and Ecker 2017), or providing tips to spot misinformation (Guess et al. 2020; Hameleers 2020).

This literature on misinformation has largely focused on developed Western countries with public social media applications such as Facebook and Twitter, where platforms have the capacity to correct misinformation by adding a disputed or warning tags on a dubious messages. These solutions may be insufficient for misinformation in developing countries, where populations have different vulnerabilities, but also where misinformation spreads largely on private and enclosed platforms.

One such platform is WhatsApp, a chat-based messaging application that is now used by over 2 billion active users in the world. On WhatsApp, messages are protected by end-to-end encryption, and such privacy renders impossible platform-based solutions such as adding disputed or warning tags to messages. As a result, observers note that WhatsApp has become a black hole for misinformation, and rumors that spread over the platform have been linked to electoral violence and social conflict in several countries (Bowles, Larreguy, and Liu 2020; Cheeseman et al. 2020; Perrigo 2019). WhatsApp users disproportionately reside in developing countries, where the mechanisms of information processing, belief in partisan content, and motivated reasoning might differ from findings in the Western context. Put together, these factors highlight key gaps in the literature on the correction of misinformation: little is know about correcting misinformation in developing countries, and WhatsApp (or discussion apps) as a medium pose technical challenges to correction.

We address these issues in the context of India, the world's largest user base for WhatsApp. In 2019, the Reuters Institute India Digital News Report found that WhatsApp was the most widely used social network platform in India: of their respondents, 82% used the application and 52% reported getting news on it, far higher numbers than most markets in Europe and North America (Aneez et al. 2019). Moreover, 40% of WhatsApp news users said they forwarded a news story in the past week, an important dynamic given the ability of the platform to reach a large number of users through its forwarding features. Overall, India is a much more mobile-first online news environment than even other developing markets like Brazil and Turkey, let alone Western contexts like the United States or Germany.

But an important reason contributing to the WhatsApp's popularity is also at the heart of the misinformation problem: WhatsApp messages are private and protected by encryption. This means that no one, including the app developers and owners themselves, have access to see, read, filter, and analyze text messages. A WhatsApp group can

exist without anyone outside the group knowing of its existence, who its members are, or what is being shared. This feature prevents surveillance by design, making WhatsApp akin to a black hole of misinformation (Ponniah 2019). In the context of India, political, sectarian and majoritarian causes have embraced this technology to diffuse falsehoods: misinformation campaigns are often run by political parties with nationwide "cyber-armies", targeting political opponents, religious minorities and dissenting individuals (Poonam and Bansal 2019). The consequences of such rumors in India have been as extreme as political unrest and violence, underscoring the dual stresses of WhatsApp and polarization on an electorate that is relatively new to the internet.

Social Corrections: A Fix For WhatsApp Misinformation?

Given the private nature of WhatsApp group chats, where misinformation cannot be corrected by the platform, any solution to misinformation must involve users correcting each other. A few studies have explored the effect of such social corrections 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. However, little is known about whether such social corrections work in the context of WhatsApp, where they cannot be compared to platform corrections and where dissemination takes place in homophilic and private group chats.

The literature on source credibility helps explain why social 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 trust-

worthiness are components of source credibility, the latter is found to be more effective in persuasion than the former (Swire and Ecker 2018). Thus, arguably, peers 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, or 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 (Housholder and LaMarre 2014).

Social media networks, and particularly group chats on platforms such as WhatsApp, are built on the basis of homophily, similarity, and shared interests. In the case of WhatsApp, this manifests in the form of groups created for specific purposes: political causes, parent-teacher associations, groups dealing with resident welfare issues. Users on a WhatsApp group chat may not know each other personally, as group ”admins” have the capacity regulate entry and exit in a group. This may especially be the case for political WhatsApp groups, where party volunteers may go door to door to add constituents (who may or may not know each other) to a WhatsApp group. Thus while the sender of a single message may remain personally unknown, group members are likely bound by shared solidarity to the group cause, and this homophily might make information from unknown sources appear highly credible (Davies 2020).

Thus social corrections in these contexts are likely to reduce the uptake of misinformation. We build on the literature on social corrections in Western contexts and seek to test whether they are equally effective in the WhatsApp ecosystem. Accordingly, we hypothesize the following:

Hypothesis 1: Exposure to corrective messages emanating from peers will reduce the perceived accuracy of misinformation, relative to a no correction condition.

Does Partisan Motivated Reasoning Affect Corrections?

The empirical literature on misinformation demonstrates that the success of fact-checking interventions is a function of individuals' preexisting beliefs. In this regard, a primary factor influencing the efficacy of corrections is partisan 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 conclusion) 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).

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, however, that partisan motivated reasoning will affect corrections is not a foregone conclusion. On the one hand, partisan affiliations in India have been shown to be traditionally been weaker and less stable, and sometimes forming for non-ideological reasons (Chandra 2004; Bussell 2019). On the other hand, reports show that misinformation is largely political on WhatsApp, with groups formed by the ruling Bharatiya Janta Party (BJP) often morphing into havens of misinformation and hateful rhetoric capable of inciting violence (Arun 2019; Farooq 2017). Indeed, research on WhatsApp groups demonstrates that a majority of political content comes from groups allied with the BJP (Garimella and Eckles 2020). Further, since we conduct this experiment after a contentious election, during which attachments to parties were arguably heightened (Michelitch and Utych 2018), we might expect motivated reasoning to play a role in information processing.

Despite these contrasting priors, in keeping with findings from the literature on

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

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

Hypothesis 2b: Peer corrections will be less effective when misinformation is attributed to an ideologically congenial politician (compared to when it is unattributed).

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

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

Hypothesis 4a: Peer corrections will be less effective when the slant of the story is ideologically congruent (compared to non-ideological stories).

Hypothesis 4b: Peer corrections will be more effective when the slant of the story is ideologically dissonant (compared to non-ideological stories).

The Role of Substantiation

If social corrections are effective, what kinds of corrections are *most* effective? Corrections could take on a variety of forms: they could be short or long; substantiated or unsubstantiated. If they are backed by evidence, such evidence could stem from a variety of sources.

Research in communication demonstrates that the persuasive quality of an argument is a function of whether or not it is substantiated (Stiff and Mongeau 2016): a body of scholarship suggests that the most persuasive arguments are ones that provide

evidence in the form of data or supporting materials (German et al. 2016). Such evidence has been found to produce general persuasive effects that appear stable (Reinard 1988). As applied to correcting misinformation, Nyhan and Reifler (2015) identify two distinct types of corrective messages: 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.

We apply these insights to our experiment and test whether the level of substantiation in corrections matters for reducing the uptake of misinformation. We distinguish between substantiated corrective messages, or those that are backed by an explanation, and unsubstantiated corrections which have no source or explanation and are consequently shorter. Further, we examine whether variations in the source of substantiation can affect beliefs in misinformation. Following concerns surrounding the increase of social media misinformation in India, a host of independent organizations and institutions launched initiatives to correct online misinformation. These included actors connected to the state (officials, policemen, teachers), domain experts, journalists, fact-checking organizations and social media platforms themselves. However, since the domain of fact-checking is a relatively new enterprise in the country and given that platform users can rely on several different actors to source a corrective message, whether citing a source matters for the efficacy of corrections in the Indian context is an open empirical question. Following from the literature on the persuasive role of substantiation, we thus hypothesize the following:

Hypothesis 5: Exposure to substantiated corrections will reduce the perceived accuracy of misinformation relative to unsubstantiated corrections.

Design

To test these hypotheses, we designed and fielded an online experiment in India (N $\approx$ 5,100). In our experiment, respondents were randomly assigned to one of four conditions in a between-subjects design (see Figure 1). In all conditions, respondents were shown hypothetical conversations on WhatsApp group chats. In our three treatment groups, these WhatsApp conversations included a misinformation stimulus posted by a user, followed by a correction posted by a second user.

Since our goal is to test the efficacy of different types of corrections, each treatment group varied the degree of substantiation of the corrective message, as well as its source. Respondents in our *Domain Expert* treatment read a substantiated correction in which the user posting the correction pointed to a domain expert as the source of the correction (for example, the Election Commission of India for electoral misinformation, or the Census Bureau of India for demographic misinformation). Respondents in our *Fact Checker* treatment read a substantiated correction in which the user posting the correction pointed to a verified fact-checker in India as having debunked the misinformation posted. We further randomized the specific fact-checker such that respondents read a fact-check from one of five sources: the online platform WhatsApp itself, the online platform Facebook, India's oldest print newspaper The Times of India, left-leaning third party fact-checking service AltNews, or right-leaning third party fact-checking service Vishwas News.

Respondents in our *Unsubstantiated Correction* treatment, on the other hand, read a correction that was a simple rebuttal by the second user, devoid of substantiation or a source of correction. This included a one-line simple correction (for instance saying "I don't think that's true, bro!") in response to the misinformation stimulus. We compare these treatment conditions to our control group, where respondents read the misinformation stimulus without a correction. The full text of each experimental manipulation, along with samples of the experimental stimuli, are included in Online Appendix C.

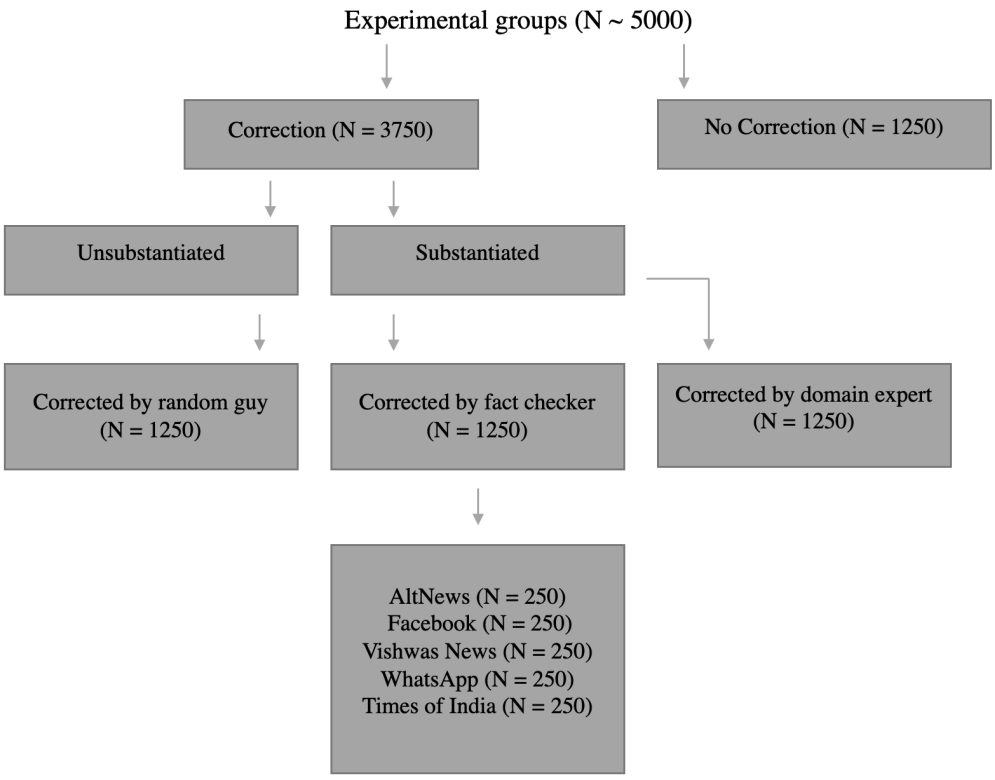

Figure 1: Experimental Design

Apart from experimentally varying the presence and sophistication of the correction, we test for partisan motivated reasoning by including information about the news outlet / source reporting the false story, as well as the political identity of politician amplifying the misinformation (where applicable). Qualitative evidence 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 (Garimella and Eckles 2020). To reflect this, we include a majority of stories credibly emanating from BJP politicians. In some cases, misinformation stimuli are shown to be emanating from the BJP's national-level competitor, the Indian National Congress (INC), while in other cases they are not attributed to any politician.

Further, we also vary the media outlet reporting the story to include three possible sources: India TV (a relatively pro-BJP private news channel), NDTV Hindi (a relatively

anti-BJP private channel), and DD News (public channel). Respondents were equally likely to be randomized into one of the four correction conditions listed in Fig.1. Further, within each treatment condition respondents had an equal probability of being assigned to each of the possible combinations of media outlets x politicians listed above.[2]

Outcomes

We measure the effect of these corrective treatments on the perceived accuracy of a series of nine stories that went viral during the 2019 elections in India. Each experimental prompt focused on one rumor that we subsequently asked respondents to evaluate. After reading the WhatsApp conversation, 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)

To ensure respondents became familiar with the format of the WhatsApp conversations and to build up the expectation that some of the stories included in the outcome measure were true, all respondents began by rating the perceived accuracy of a true story (that is, one featuring a verifiable and uncontroversial claim). 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. These stories are listed in Table 1.

[2]A small proportion of respondents (3% of the sample) 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 Appendix Table J.1, we detect no statistical differences in the overall rate of belief in all false stories when comparing our pure control condition to control condition with no correction.

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. Each of the stories selected for the final experiment were strongly believed or believed by at least of 25% of the pretest sample, with some of these statements being believed by a large majority of respondents. Data from this pretest is presented in Appendix I.

The final selection of stories was the product of several constraints and choices. To avoid prompting respondents to systematically reject the veracity of rumors, we included some true stories. But simultaneously, our goal was to maximize 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).

Our design 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, 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 slightly varied the specific text of the messages in each screenshot to ensure realism. Online Appendix C gives a full list of treatments for each rumor used in our experiment and provides a sample of WhatsApp screenshots shown to treatment group respondents.

Sample

Participants in this study were Hindi speakers recruited through Facebook. The ad used to recruit respondents is presented in Online Appendix B. To be eligible to participate, respondents were required to be adult residents of India who used WhatsApp. While we recruited over 5100 participants, the actual N presented in our analyses varies slightly for each dependent variable story (+ or - 1%), as we include observations from respondents who exit the survey before reading all 9 screenshots.[3] The experiment was conducted entirely in Hindi. Sample characteristics of our respondents are in Appendix M.

Results

This section begins with descriptive analyses that demonstrate the extent of belief in misinformation. Figure 2 lists the 7 false stories used in the dependent variable measure in this study. This figure plots the share of respondents in the sample who believed each story to be true. Our findings demonstrate the high salience of false stories in the Indian context. Despite the fact that 75% of all screenshots (across conditions) contained a correction, 6 of the 7 false rumors were rated as accurate or somewhat accurate by

[3]Our results are robust when we limit the sample to those who finished the entire survey.

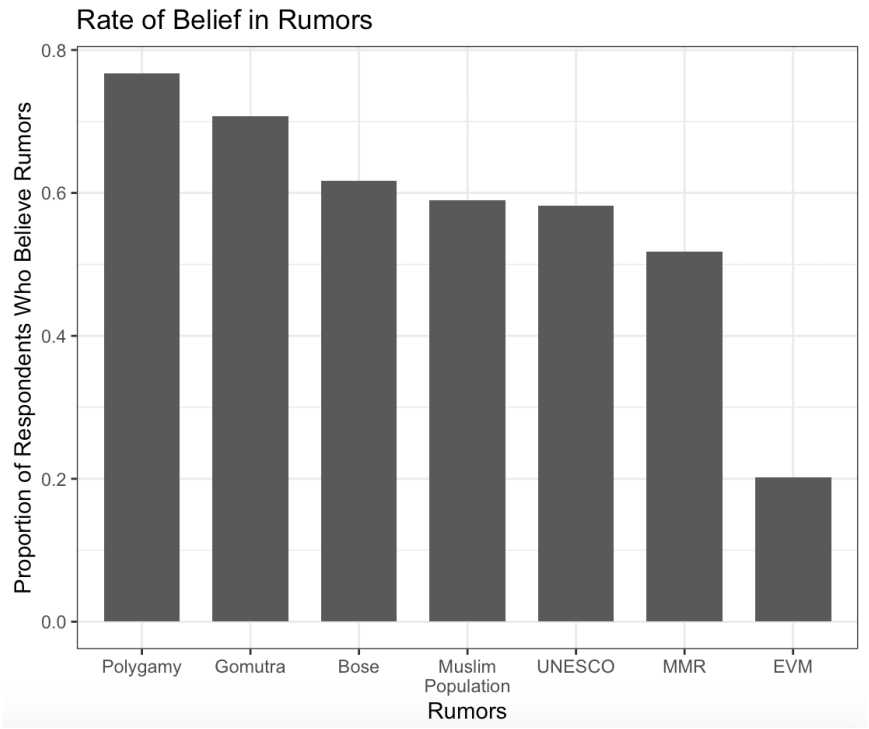

Figure 2: Overall rate of belief in misinformation across experimental conditions

over 50% of the sample, with the top two prevalent rumors believed by over 75% of the sample.

Do Corrective Messages Matter?

Do corrections impact these high levels of belief? We first present results for Hypothesis 1, which tests whether exposure to any correction reduces the perceived accuracy of misinformation. To test Hypothesis 1, we pool together all the different types of social corrections such that the primary comparison of interest is between having received a correction (of any kind) and not having received one. This comparison is expressed in Equation 0.1:

$$Belief\,Accuracy_i = \alpha + \beta_1 AnyCorrection_i + \epsilon_i \tag{0.1}$$

In the equation, the *AnyCorrection* variable represents pooled assignment all cor-

rection 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 greater perceived accuracy. We estimate a separate bivariate OLS model for each of the seven false stories in our experiment, represented by the seven columns in Table 2.[4]

Our main result demonstrates that corrections are effective at reducing beliefs about the accuracy of false stories. Exposure to social corrections of any kind significantly reduces the likelihood that respondents report false rumors to be accurate, relative to not receiving any correction.

We do not obtain a significant result for only one (out of 7) 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 treatment 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.[5] Thus our results show that social corrections, the only suitable techniques for private online spaces, significantly reduce overall rates of beliefs in patently false rumors circulating on WhatsApp.

[4]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 Appendix Table J.1 that there is no difference between the control and pure control conditions.

[5]The size of the effect across rumors does not appear related to the prior salience of these rumors in our sampled population. As shown in Appendix 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.

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

This result holds in the presence of added control variables. In Appendix Table K.1, we control for the media source reporting the story as well as the politician it is attributed to, two factors likely to impact beliefs. We find that our results are robust to these controls: regardless of the the media source or politician being congruent or dissonant, results hold. Our experimental corrections significantly improve the perceived accuracy of beliefs, and the magnitude of these effects remains relatively unchanged.

To our knowledge, this is the first experiment to test the effect of any sort of correction strategy to temper the uptake of misinformation on WhatsApp, and the first to systematically test the effect of any corrective strategy in India. Our finding that witnessing corrections posted by peers can subsequently reduce the perceived accuracy of false stories is a hopeful one, given the cost of misinformation in polarized contexts like India. We are able to demonstrate that social corrections can be a useful tool to counter misinformation in this sample, at least in the short run.

Since these results emerge from a controlled, experimental context, we cannot entirely exclude the possibility that the intensity of corrections would differ in real life. However, these findings indicate that corrections by an unidentified peer consistently

produce reductions in misinformed beliefs despite any lack of incentive to answer correctly. Hence, these results suggest that credible corrections on actual chats built around the principle of homophily may be as or more impactful.

Motivated Reasoning and Social Corrections

To what extent are the corrective effects obtained above affected by motivated reasoning in the Indian context? To answer this question, we look at motivated reasoning in three ways. First, we examine whether the partisan identity of the politician making the false claim affects belief in the story. In India, as is often the case in several other contexts, misinformation can emanate from elites who amplify false stories for political gain. Misinformation coming from the top can often have a stronger effect on respondents who are ideologically inclined with elites making false claims, relative to misinformation coming from non-elites (Van Duyn and Collier 2019). Hence it stands to reason that congruence between respondents' partisan identity and the identity of the politician to whom a story is attributed can impact beliefs.

To test Hypotheses 2a and 2b, we limit our analyses to the subset of stories that are clearly partisan in nature (Rumors 3, 4, 6, 8,and 9) and code whether the rumor was attributed in the experimental prompt to a copartisan or outpartisan politician. We code a politician as copartisan or outpartisan as a function of the respondent's self-reported partisan inclination towards the ruling party, the BJP, relying on the respondent's expressed closeness to this party. Concretely, a BJP politician is deemed copartisan if the respondent describes themselves as close or very close to the party, and outpartisan if the respondent describes themselves as far or very far from it. By contrast, a INC politician is deemed copartisan if the respondent describes themselves as far or very far from the BJP and outpartisan if the respondent describes themselves as close or very close to the BJP.

Second, we look at the effect that congruent or dissonant media outlets report-

ing a story can have on beliefs (Hypotheses 3a and 3b).To test Hypotheses 3a and 3b, we code a media outlet as congruent 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 congruent 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 congruent when the respondent reports feeling far or very far to the BJP.

Finally, we look at the degree of congruence of the story slant itself with respondents' beliefs. To what extent does the efficacy of the treatment depend on whether the content of the false story was congruent with respondents' political beliefs (Hypotheses 4a and 4b)? To test Hypotheses 4a and 4b, we again limit our analyses to the subset of rumors that are clearly political (rumors 3, 4, 6, 8,and 9). We code rumors as congruent or dissonant ex-ante as a function of participants own ideological inclinations and as a function of our observations of the 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 congruent 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.[6]

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 or congruence of the source/media outlet/rumor. In Tables 3 and 4, we test whether the effect of the correction is a function of the slant of the story itself. While

[6]Results in Appendix Tables G.1 and G.2 underscore this coding choice: we show that rumor congruence 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.

Table 3 looks at whether corrections are less effective for ideologically congruent stories, Table 4 looks are whether corrections are more effective for ideologically dissonant stories (Hypotheses 4a and 4b).

Table 3: Effect of Correction * Congruent 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)
CongruentClaim	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
CongruentClaim	(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

Across both tables, the interaction between the treatment (any correction) and the slant of the story corrected produces a null result. The results from these tests thus point to a striking conclusion: the effect of corrections is *not* limited by the slant of the story. Similar results emerge when we look at motivated reasoning in two other ways: the politician to whom a given story is attributed, and the news outlet reporting the story. We find that interacting the correction with the identity of a given politician or media outlet also does not reduce or change the effect of corrections. Results from these tests are reported in Appendix Tables D.1 and D.2 (effect of the identity of the politician) and Tables E.1 and E.2 (effect of media outlet).

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

	Dependent variable: Belief in Rumor				
	MuslimPop	Polygamy	Gomutra	EVM	UNESCO
	(1)	(2)	(3)	(4)	(5)
AnyCorrection	−0.127***	−0.195***	−0.088**	−0.004	−0.032
	(0.045)	(0.036)	(0.039)	(0.049)	(0.040)
DissonantClaim	−0.206***	−0.188***	−0.267***	−0.608***	−0.267***
	(0.069)	(0.055)	(0.058)	(0.053)	(0.060)
AnyCorrection*	0.062	0.016	−0.060	−0.022	−0.058
DissonantClaim	(0.078)	(0.064)	(0.069)	(0.062)	(0.070)
Constant	2.816***	3.334***	3.097***	2.022***	3.058***
	(0.040)	(0.031)	(0.033)	(0.042)	(0.035)
Observations	5,104	5,103	5,099	5,136	5,099
R^2	0.006	0.015	0.021	0.090	0.019
Adjusted R^2	0.006	0.015	0.020	0.089	0.019
Res. Std. Er.	1.093 (df = 5100)	0.942 (df = 5099)	1.050 (df = 5095)	0.968 (df = 5132)	1.051 (df = 5095)
F Statistic	10.658***	26.475***	36.056***	168.534***	33.051***
Note:					*p<0.1; **p<0.05; ***p<0.01

Taken together, these results demonstrate that the effectiveness of corrections persists despite partisan ties, and that perceived accuracy of rumors is *not* a function of the source or slant of a rumor. This suggests that partisan motivated reasoning plays a comparatively less important role in India, breaking with results frequently obtained in the American context (Nyhan and Reifler 2010). We find that motivated reasoning, however measured, does not condition the effect of corrections.

We also test whether partisans of the ruling party (the BJP) are susceptible to partisan motivated reasoning. Since there is a supply side bias in political motivation in India (with misinformation often emanating from BJP-sympathetic sources) and since we code party identity as feelings towards the BJP, it stands to reason that BJP party identity is more consolidated in our sample (those not supporting the BJP may support a host of other national and regional parties in the country). We find that even amongst this subgroup of arguably stronger and more consolidated partisans, no motivated reasoning effect exists (see Appendix Table F.1). This absence of effects persists when we run a

second series of tests relying on reported voting decisions in the 2019 elections instead of measuring respondents' closeness to the BJP: participants who 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. We explore this result further in the discussion below.

Are Substantiated Corrections More Effective?

Our main effect in this paper demonstrates that receiving any correction (relative to control) can reduce the perceived accuracy of misinformation. While this analysis pools together all corrections, we now examine which types of corrections are most effective. Particularly, we compare substantiated to unsubstantiated messages, and determine whether the source of substantiation plays a role in persuasion.

In Table 5, we evaluate the effect of different types of corrections on the level of belief in each of the rumors, compared to the control condition. The first row of results in Table 5 represents corrections without substantiation, in which a peer in the group chat expresses skepticism with the story but does not explain the source of the skepticism (merely stating a version of "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.

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 we reinforce in the graphical representation of these results (see Online Appendix L). This implies that the sophistication of social corrections matters very little: the unsourced correction is often as effective as the longer and more clearly sourced corrections in this design. An unidentified participant merely expressing incredulity about a rumor is therefore

Table 5: Effect of Correction Source

	Dependent variable: Belief in Rumor						
	MuslimPop	Polygamy	MMR	Gomutra	EVM	UNESCO	Bose
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Unsourced Correction	-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)
Domain 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 News	-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)
Times of India	-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

as likely to reduce belief in a falsehood as a respondent engaging in a more careful 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, 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. 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. On the contrary, merely indicating that a message might be false, relative to sourcing that indication, appears to be enough to move beliefs. Overall, we thus find that the content of social corrections counts less than the mere presence of one.

Discussion

One of the most striking findings to emerge from this study demonstrates that exposure to a corrective message posted by an unidentified peer on a WhatsApp chat – regardless of the source or substantiation of that message – works to reduce beliefs in misinformation. 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 could achieve a much larger effect.

These results additionally show that corrective effects can be achieved in contexts very different from the United States. India has lower rates of formal education and digital literacy, as well as users who are new to the Internet. These factors likely imply that news received via the Internet might automatically have more value, given the

unfamiliarity and fascination the medium inspires. While this may lead misinformation to be more easily believed, the same may apply to corrections, which may more easily become effective in such contexts.

On the other hand, and in stark contrast with the U.S.-centered literature on corrections, our results suggest that different forces drive the corrective process in developing countries such as India. The absence of partisan motivated reasoning in our results - despite our best efforts to detect any - is especially striking. While we can only speculate as to the causes of this divergence, one explanation may lie in the nature of partisanship in India. Despite much report of polarization and of the increasingly common transformation of partisanship into a social identity (Chhibber and Verma 2018), India is a country that has traditionally had weaker partisan ties, and where politics is thought to be more clientelism-based or ethnicity-based rather than programmatic (Auerbach et al. 2021). This relative weakness of partisanship - at least in the American sense of the term - may imply that motivated reasoning would *not* constitute as big an obstacle to correcting beliefs, or more likely in our view, that partisanship may not be the basis for motivated reasoning in this context (it may instead exist in another, non-partisan form, for instance, along the lines of religious identities).

Additionally, we demonstrate that respondents react similarly to substantiated and unsubstantiated corrections, and that the presence of a correction, rather than its source, appears sufficient to change beliefs. This finding underscores the presence of norms of solidarity that guide behavior in homophilic WhatsApp group chats, where source credibility may be derived from membership to the group itself. In other words, the subset of respondents whose beliefs were amenable to correction did correct their beliefs no matter the type of correction, since they perceived the existence of a strong social norm. While we lack empirical evidence for this mechanism, this line of argument echoes recent findings in the literature on social media - in a variety of contexts - suggesting that social norms powerfully constrain online behaviors and beliefs. Both

Munger (2017) and Siegel and Badaan (2020) show that online hatred and harassment can be significantly reduced by simple nudges, especially if these come from ingroups. Further, Groenendyk and Krupnikov (2020) demonstrate that information processing is motivated by the goals made salient in a given context. While political contexts may invoke conflict, WhatsApp groups formed around shared causes may push respondents to seek consensus around a common goal. Following this line of argument, our respondents appear to have adopted the dominant (or in our case, the loudest, last and more visible) group belief, regardless of its sophistication or substantiation.

Our own intuitions about WhatsApp users in India suggest that this logic may be particularly central to the behavior of the users of private discussion apps such as WhatsApp. In focus groups and interviews we carried in recent months in India to better understand online behavior, conformity pressure to group norms emerges as a leading driver of behavior. Many users in closed groups may feel pressure not to believe or say something that runs against the dominant view that anchor the group's identity (Davies 2020). Thus testing the precise mechanisms for belief in misinformation and its correction will be an important question for future research in this context.

Conclusion

Misinformation campaigns have the capacity to affect opinions and elections across the world. Purveyors and victims of misinformation and hyper-partisan messaging are no more just individuals with low digital literacy skills, people who are uninformed, Internet scammers, or Russian trolls. A global rise in polarization has meant that the creators and contributors of misinformation include party workers, stakeholders and politicians themselves. As the world moves to deal with the COVID-19 crisis, we are engulfed in a new deluge of misinformation in hyper-partisan and polarized environments, where traditionally non-political issues are also deeply politicized. The rise of polarization amidst

a global pandemic underscores the need to identify robust strategies to counter the pernicious effects of misinformation, especially in societies where it is spread on private platforms and where the stakes are as high as violence.

In this paper, we present new evidence on corrective strategies to temper the uptake of misinformation in developing contexts that use WhatsApp, with evidence from an experiment in India. We demonstrate that peers correcting each other on encypted group chat networks can be an effective tool to reduce perceived accuracy of misinformation. Relative to a no correction condition, receiving any correction to a misinformation stimulus significantly reduces beliefs in false stories. Further, we find that this effect is not limited by partisan motivated reasoning of any kind. Finally, we demonstrate that the source or sophistication of these corrective messages matters less relative to their presence, as unsubstantiated messages achieve an effect comparable to those that are backed by evidence.

Despite our promising findings on the effects of corrections, we now consider some limitations of the study and future avenues for research. A first limitation derives from the fact that we measure our dependent variable in close proximity to the treatment, and thus cannot speak to the durability of these effects. Future studies should consider a longer gap between treatment and outcomes to measure whether such effects decay over time. Next, our design over-sampled false news stories in the outcome measure. Future studies can vary the ratio of true vs. false stories to examine the effects on the efficacy of the treatment. Finally, a challenging prospect for future research is to be able to examine the effect of corrections in a more naturalistic setting, outside of a survey or online experiment. While the private nature of WhatsApp groups makes this difficult, measuring the impact of misinformation and solutions to counter it within the ecosystem of groups that individuals are a part of will allow us to ascertain the true impact of group solidarity, conformity pressures, and ingroup norms.

The policy implications of our findings are mixed. Our results constitute the first

test of corrective treatments for misinformation in India and confirm that peer-to-peer fact-checking can improve information processing. 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: 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 a chat. Platforms may help with this 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. 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. 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, especially if users are easily prone to conformity pressures. 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. While this raises ethical challenges, future research could analyse whether a peer posting a factually incorrect correction may have a similar effect on belief change.

Beyond chat apps and other 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 settings of low education, accelerating Internet, and private online spaces. 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

Aneez, Zeenab, Ahmed T Neyazi, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. 2019. "India Digital News Report." *Reuters Institute for the Study of Journalism*. https://bit.ly/3x3Gpnl.

Ara, Ismat. 2020. "Tear Them Apart: How Hindutva WhatsApp Group Demanded Murder, Rape of Muslims in Delhi Riots." *The Wire*.

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

Auerbach, Adam Michael, Jennifer Bussell, Simon Chauchard, Francesca R. Jensenius, Gareth Nellis, Mark Schneider, Neelanjan Sircar, Pavithra Suryanarayan, Tariq Thachil, Milan Vaishnav, and et al. 2021. "Rethinking the Study of Electoral Politics in the Developing World: Reflections on the Indian Case." *Perspectives on Politics*: 1–15.

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.

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

Bussell, Jennifer. 2019. *Clients and constituents: Political Responsiveness in Patronage Democracies*. New York: Oxford University Press.

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

Chandra, Kanchan. 2004. *Why Ethnic Parties Succeed: Patronage and Ethnic Headcounts in India*. New York: Cambridge University Press.

Cheeseman, Nic, Jonathan Fisher, Idayat Hassan, and Jamie Hitchen. 2020. "Social Media Disruption: Nigeria's WhatsApp Politics." *Journal of Democracy* 31 (3): 145–159.

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

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

Davies, William. 2020. "What's Wrong With WhatsApp." *The Guardian*. July 2, 2020. https://www.theguardian.com/technology/2020/jul/02/whatsapp-groups-conspiracy-theories-disinformation-democracy.

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.

Garimella, Kiran, and Dean Eckles. 2020. "Images and misinformation in political groups: evidence from WhatsApp in India." *Harvard Kennedy School Misinformation Review*.

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.

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

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

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

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.

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.

Munger, Kevin. 2017. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." *Political Behavior* 39 (3): 629–649.

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/.

Ponniah, Kevin. 2019. "WhatsApp: The 'black hole' of fake news in India's election." *BBC News*. April 6, 2019. https://www.bbc.com/news/world-asia-india-47797151.

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

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.

Siegel, Alexandra A, and Vivienne Badaan. 2020. "# No2Sectarianism: Experimental approaches to reducing sectarian hate speech online." *American Political Science Review* 114 (3): 837–855.

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

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.

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

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.

Van Duyn, Emily, and Jessica Collier. 2019. "Priming and fake news: The effects of elite discourse on evaluations of news media." *Mass Communication and Society* 22 (1): 29–48.

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.

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

Online Appendix

"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

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 weblsite.. This BJP guy said UNESCO declared PM Modi best Prime Minister in 2016.	Just saw this on NDTV Hindi/Doordarshan/India TV weblsite.. 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

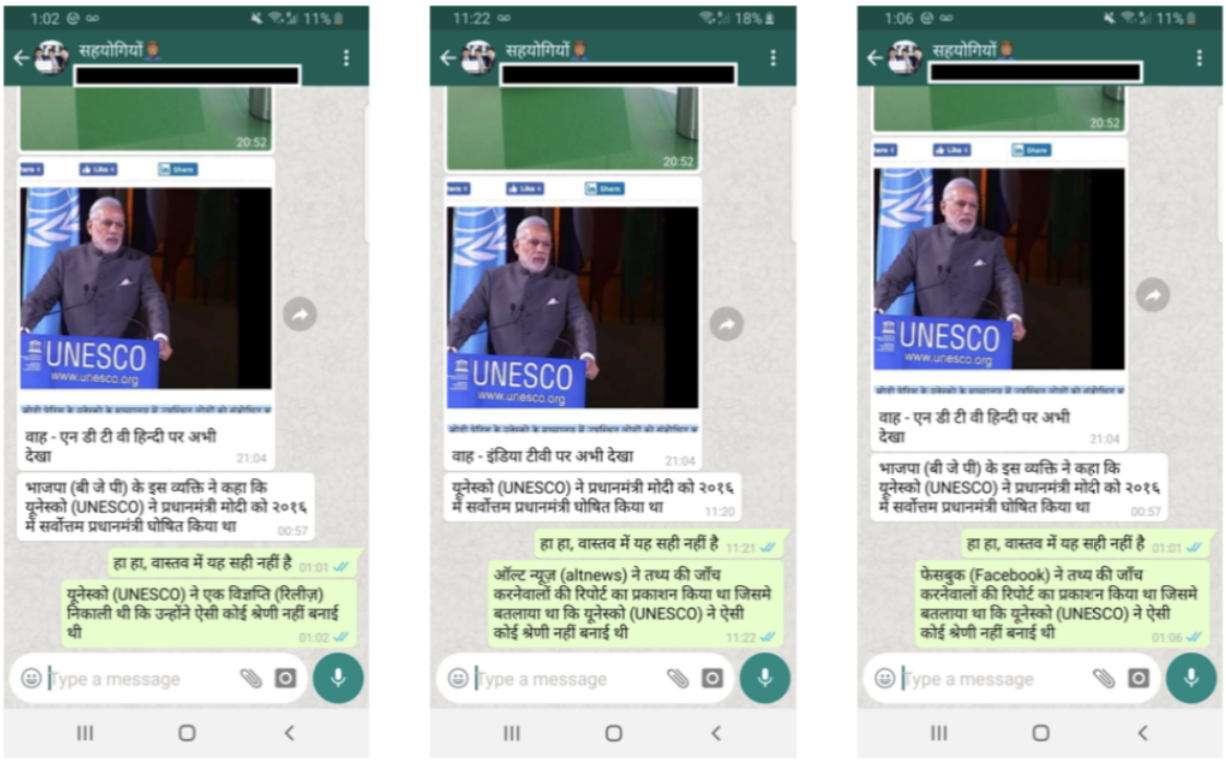

Figure C.2: Sample WhatsApp Conversation

D Tests For 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).

$$Belief\,Accuracy_i = \alpha + \beta_1 AnyCorrection_i + \beta_2 CongenialPol_i + \beta_3 AnyCorrection * CongenialPol_i + \epsilon_i \quad \text{(D.2)}$$

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 Tests For 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 Media	-0.224^{*}	0.230^{**}	-0.078	0.080	-0.163	0.036	0.186
	(0.126)	(0.113)	(0.121)	(0.108)	(0.111)	(0.151)	(0.114)
Dissonant Media	-0.132	0.061	-0.050	-0.184	0.051	0.094	-0.048
	(0.121)	(0.108)	(0.121)	(0.116)	(0.118)	(0.143)	(0.116)
AnyCorrection* CongenialMedia	0.271^{**}	-0.194	-0.049	-0.056	0.080	-0.067	-0.174
	(0.135)	(0.122)	(0.131)	(0.119)	(0.121)	(0.162)	(0.125)
AnyCorrection* DissonantMedia	0.219^{*}	-0.134	-0.067	0.272^{**}	-0.134	-0.092	0.091
	(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.

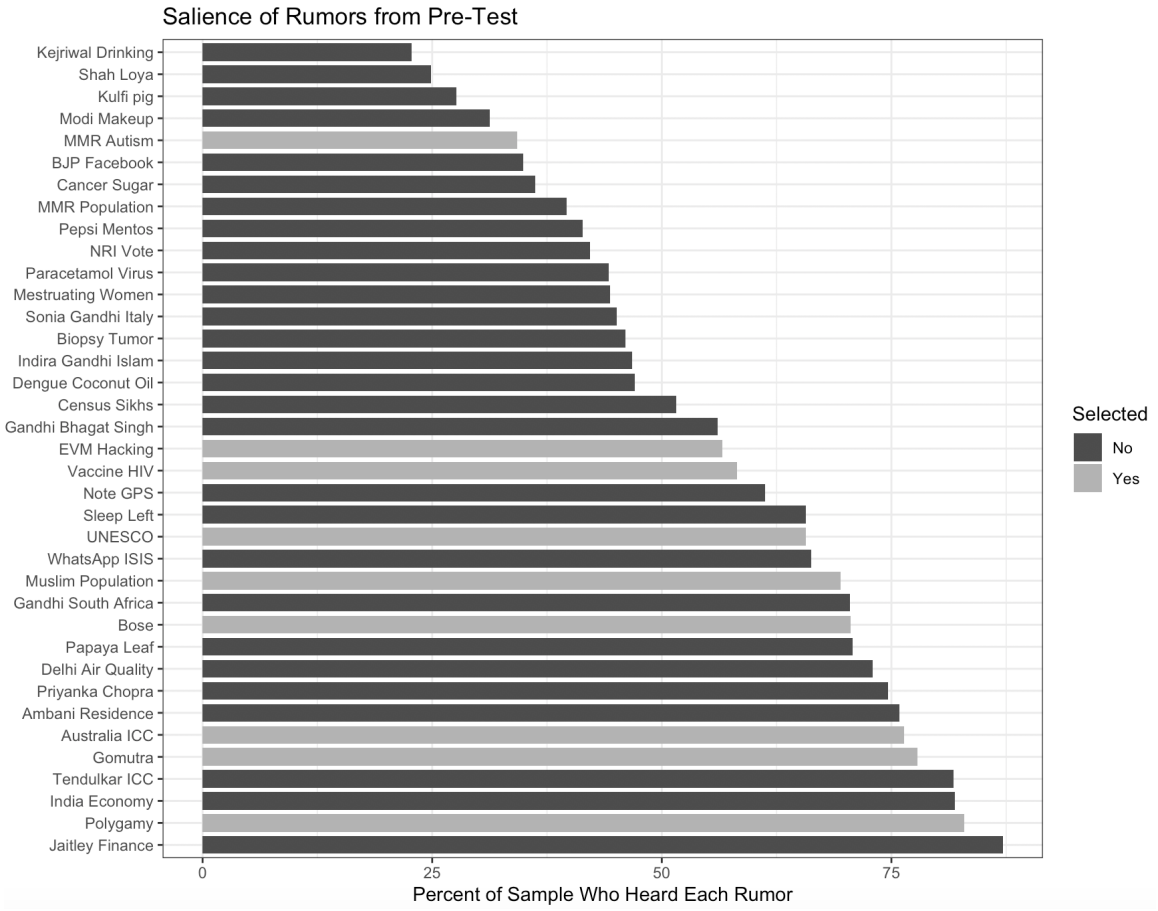

Figure I.1: Salience of Pretest Rumors

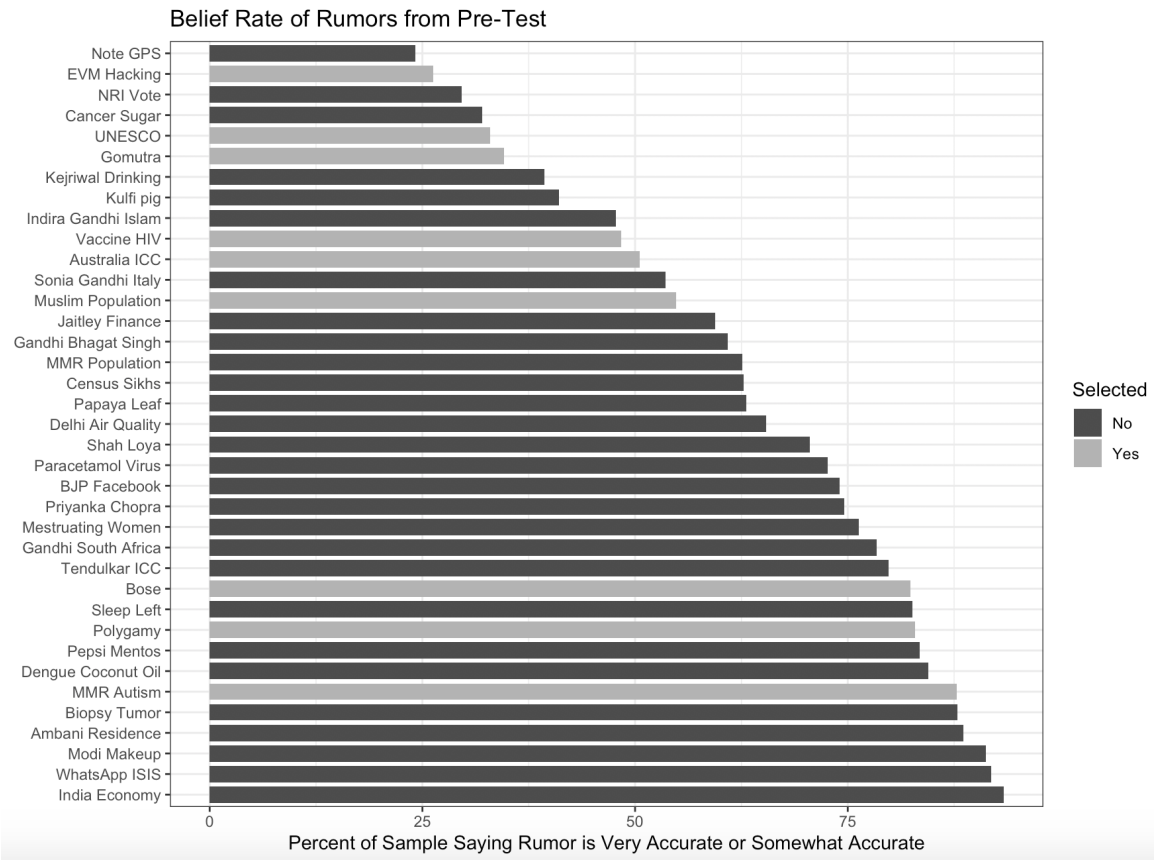

Figure I.2: Belief in Pretest Rumors

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

K Main Effect With Controls

Table K.1: 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)
Congruent 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 Politician	-0.022	-0.008		0.091^{*}	0.434^{***}	0.091	
	(0.051)	(0.046)		(0.050)	(0.061)	(0.060)	
Outpartisan Politician	-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

L Effect of Correction Source

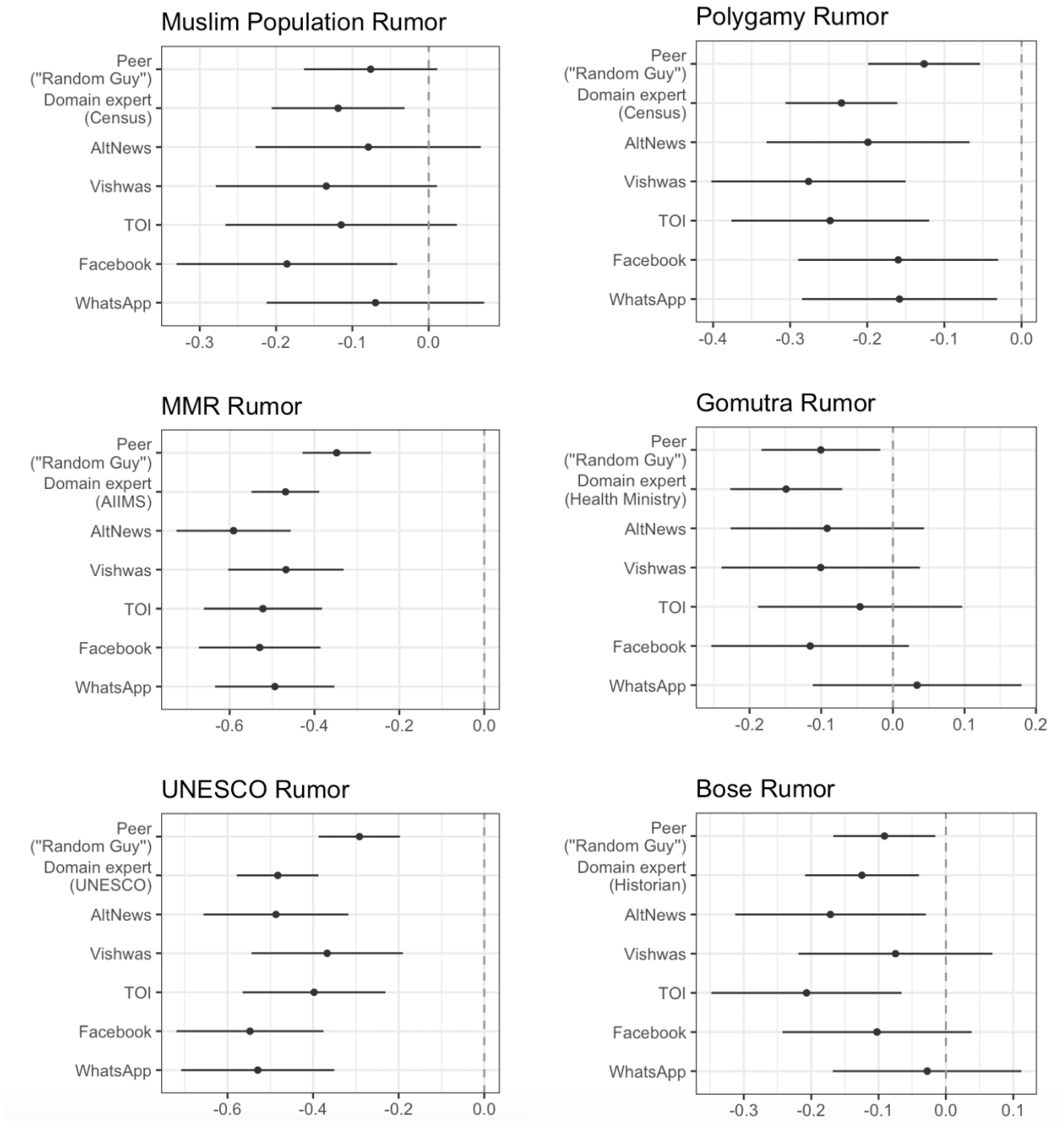

Figure L.1: Effect of Correction Source

M Sample Characteristics

Table M.1: Summary Statistics of Key Variables

Statistic	N	Mean	St. Dev.	Min	Median	Max
Age	4,948	29.68	9.43	18	27	76
Male	5,136	0.86	0.34	0	1	1
Education	5,136	6.93	0.97	1	7	8
Hindu	5,136	0.87	0.33	0	1	1
Upper Caste (General)	5,136	0.57	0.49	0	1	1
SC / ST	5,136	0.13	0.34	0	0	1
BJP Partisan	5,136	0.65	0.47	0	1	1
Facebook Use Frequency	5,136	5.40	0.99	1	6	6
WhatsApp Use Frequency	5,136	5.65	0.83	1	6	6

N Analysis Plan

This appendix contains the analysis plan for this study as registered with EGAP.

N.1 Hypotheses

We test the following hypotheses:

H1: Exposure to corrective information on WhatsApp will reduce the perceived accuracy of the targeted claim relative to a pure control condition.

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

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

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

H3b: WhatsApp corrections will be less effective when the targeted claim originates from a congenial media outlet (compared to an unattributed or neutral outlet).

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

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

Besides, in line with the aforementioned main objective of our study, we also investigate the following research question on which we have weaker theoretical priors:

RQ1: Which types of corrections are most efficient at reducing misinformation? Are professional fact checking outlets more effective at combating misinformation than domain experts and/or platforms (i.e., WhatsApp)?

N.2 Experimental Design

We test our hypotheses with a survey experiment conducted on a sample of Hindi speakers recruited from Facebook, a common vector of misinformation in India. The survey consists of two sections: a pre-treatment section (hereafter "section 1") and an experimental section (hereafter "section 2"). Section 1 includes questions about respondents' demographics and pre-treatment covariates: political attitudes, conspiratorial predispositions, and trust in various political and health institutions.

Section 2 includes our experimental treatments and outcome measures. The experiment asks respondents to read and evaluate WhatsApp threads about political, health, and social claims that have circulated recently in India. Each thread focuses on one factual claim that we subsequently ask respondents to evaluate. In order to ensure respondents were used to the format of the prompts and build up the expectation that some of the claims included in the experiment were "true", all respondents began Section 2 by rating the perceived accuracy of a true claim involving the number of Australia's victories in the cricket world cup. The order of the remaining 8 claims was randomized. In total, all respondents evaluated 9 claims, 7 of which are false and 2 of which are true.

For each claim, respondents are randomized into one of several treatment conditions or, for a very small percentage of them (3%), to a pure control condition in which they do not read any thread. Respondents who are assigned to read a thread (all except those assigned to the pure control), will have equal probability of being assigned to each of the possible combinations of experimental treatments listed below.

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

1. NDTV Hindi (anti-BJP private channel).

2. India TV (clearly pro-BJP private channel).

3. Doordarshan News (public channel - hence presumably pro-government and pro-BJP).

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

1. BJP party politician.
2. INC party (anti-BJP) politician.
3. No attribution.

Experimental factor 3: presence/source of corrective message. Respondents were equally likely to be assigned to one of 4 possible conditions here, though the fourth condition was further subdivided in five, as shown below.

1. no correction.
2. unattributed non-expert correction (aka. "random guy" correction).
3. Authority/domain expert correction (i.e. correction by an expert on the claim being discussed, for instance a medical professional re. Health claims)
4. Correction by a specialized fact-checking service, such as: Altnews – in this case, respondents are told that this specialized fact-checking organization with left-leaning politics has fact-checked the claim and found it to be erroneous. Vishwasnews in this case, respondents are told that this specialized fact-checking organization with right-leaning politics has fact-checked the claim and found it to be erroneous. The Times of India – in this case, respondents are told that the country's best known/oldest newspaper has fact-checked the claim and found it to be erroneous. Facebook – in this case, respondents are told that the online platform itself has fact-checked the claim and found it to be erroneous. WhatsApp – in this case, respondents are told that the online platform itself has fact-checked the claim and found it to be erroneous.

To increase realism, we exclude highly unrealistic manipulations (e.g., voter fraud allegations attributed to ruling party politicians) and tailor domain expert corrections to each rumor (e.g., we attribute expert corrections of voter fraud rumors to the Election Commission of India). The spreadsheet appended to this pre-registration gives a full list of treatments for each rumor used in our experiment. This spreadsheet is available as a publicly viewable Google Document at this link: https://bit.ly/3mXH3y4

After reading the thread (or not, in the case of pure control respondents), respondents answer a single outcome question about their belief in the claim (shown immediately below).

Eligibility and exclusion criteria for participants: Hindi speaking Facebook users located in India (age 18+) Our primary outcome measure is belief in the central claim contained in each WhatsApp thread. (All questions appeared in Hindi. We provide the English versions here.) The format of this question is as follows:

[Thread]

How accurate is the following statement?

[Statement]

-very accurate

-somewhat accurate

-not very accurate

-not at all accurate

The wording of this factual outcome measure for each claim is included in the linked spreadsheet. The full questionnaire as administered on Qualtrics is also attached.

N.3 Statistical Analysis

We will test our hypotheses with pooled models that include all rumors that included a relevant treatment (see the design spreadsheet and the Respondents and data collection section above for a full list of rumors and treatments included in each test). These

pooled models will be estimated with OLS using robust standard errors clustered at the respondent level (i.e., since respondents are answering multiple outcome questions per above). Our factual belief outcome measures are coded on a 0-1 scale where higher values = more accurate beliefs. (If the claim mentioned in the question is true, then very accurate = 1, somewhat accurate = .66, not very accurate = .33, not at all accurate = 0. If the claim is false, then very accurate = 0, somewhat accurate = .33, not very accurate = .66, not at all accurate = 1.)

For each hypothesis below, we will also estimate separate models for each claim listed above (e.g., Australia cricket, HIV/AIDS, etc.). These models will take the same form as below but not include clustered standard errors or claim fixed effects. We plan to use Benjamini and Hochberg's (1995) method for sequentially correcting for multiple hypothesis testing.

H1: Exposure to corrective information on WhatsApp will reduce the perceived accuracy of the targeted claim.

Belief accuracy (coded 0-1 per above) = b0 + b1 Correction (1 if exposed to any correction, 0 otherwise)

H2a: WhatsApp corrections will be more effective when the targeted claim is attributed to a dissonant politician (compared to an unattributed or neutral outlet).

Belief accuracy (coded 0-1 per above) = b0 + b1 Correction (1 if exposed to any correction, 0 otherwise) + b2 DissonantlPol (1 if claim is attributed to a congenial politician, 0 otherwise) + b3 Correction*DissonantPol

H2b: WhatsApp corrections will be less effective when the targeted claim is attributed to a congenial politician (compared to an unattributed or neutral outlet).

Belief accuracy (coded 0-1 per above) = b0 + b1 Correction (1 if exposed to any correction, 0 otherwise) + b2 CongenialPol (1 if claim is attributed to a congenial politician, 0 otherwise) + b3 Correction*CongenialPol

To test H2a and H2b, we limit our analyses to the subset of rumors that are clearly

partisan in nature. The attached spreadsheet details which claims are included and which claims are coded as dissonant and congenial to members of the two parties. 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). In supplemental analyses, we will examine whether the effects of dissonance and congeniality vary by party.

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 (see relevant question on page 24/50 of pdf of instrument 1). 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.

Robustness tests for H2a and H2b:

As noted above, in supplemental analyses, we will examine whether the effects of dissonance and congeniality vary by party, and specifically whether a different reaction exist among BJP supporters vs. BJP opponents.

We also run a second series of test relying on reported voting decisions in order to establish the robustness of our findings to alternative measures of politician congeniality or dissonance. That is, instead of relying on a survey item measuring respondents' closeness to the BJP, we instead rely on the party they declare having voted for in the 2019 election.

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

H3b: WhatsApp corrections will be less effective when the targeted claim originates from a congenial media outlet (compared to an unattributed or neutral outlet).

Belief accuracy (coded 0-1 per above) = b0 + b1 Correction (1 if exposed to any correction, 0 otherwise) + b2 CongenialMedia (1 if claim is attributed to a congenial media outlet, 0 otherwise) + b3 DissonantMedia (1 if claim is attributed to a dissonant media outlet, 0 otherwise) + b4 Correction*CongenialMedia + b5 Correction*DissonantMedia

The attached spreadsheet details which media outlets are pre-determined as dissonant or congenial in our analyses. supplemental analyses, we will examine whether the effects of dissonance and congeniality vary by party.

We code a media outlet as congenial or dissonant as a function of the respondent's expressed proximity to the BJP (see relevant question on page 24/50 of pdf of instrument 1). 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.

*** H4a: WhatsApp corrections will be less effective when the targeted claim is ideologically congenial to respondents.

H4b: WhatsApp corrections will be more effective when the targeted claim is ideologically dissonant to respondents.

Belief accuracy (coded 0-1 per above) = b0 + b1 Correction (1 if exposed to any correction, 0 otherwise) + b2

To test H4a and H4b, we limit our analyses to the subset of rumors that are clearly congenial/dissonant to supporters of one of the two major parties in India: the BJP and INC. The attached spreadsheet details which claims are included and which claims are coded as dissonant and congenial to members of the two parties. 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 an anti-INC claim). In supplemental analyses, we will examine whether the effects

of dissonance and congeniality vary by party.

We code claims as congenial or dissonant ex-ante as a function of participants own ideological inclinations and as a function of our observations of these two parties' platforms. Namely, when participants self-report being "close" or "very close" to the BJP, claims number 1 (Muslim population growth), 2 (polygamy within the Muslim population), 4 (belief about the virtues of cow urine), and 8 (Modi and Unesco) are coded as congenial claims. By contrast, claim number 7 (EVMs) is coded as dissonant, while claims number 3, 5, 6 and 9 are coded as neither congenial nor dissonant. Similarly, when participants self-report being "close" or "very close" to the INC, claim number 7 is coded as congenial while claims number 1,2,4,8 are coded as dissonant and claims 3, 5, 6, and 9 are neither congenial nor dissonant.

RQ1: Which types of correction are more efficient?

To explore this question, we will run two types of models

The first one does not disentangle between the various types of fact-checking organizations listed above:

Belief accuracy (coded 0-1 per above) = b0 + b1 RandomGuy Correction (1 if exposed to a correction from an unidentified source, 0 otherwise) + b2 DomainExpert (1 if exposed to a correction from an expert source that is not a fact-checking organization, 0 otherwise) + b3 Fact Checker Correction (1 if exposed to a correction from one of the 5 types of professional fact-checking organizations listed above, 0 otherwise).

The second one does disentangle between the various types of fact-checking organizations listed above:

Belief accuracy (coded 0-1 per above) = b0 + b1 RandomGuy Correction (1 if exposed to a correction from an unidentified source, 0 otherwise) + b2 DomainExpert (1 if exposed to a correction from an expert source that is not a fact-checking organization, 0 otherwise) + b3 Altnews + b4Vishwasnews +b5 ToI + b6 Facebook + b7 WhatsApp (1 if exposed to a correction from one of these types of professional fact-checking organiza-

tions listed above, 0 otherwise).

The estimand of interest for these regressions is the difference between any two selected coefficients.

Notes:

1. We will compute and report appropriate auxiliary quantities from our models, including treatment effects by subgroup and differences in marginal effects between subgroups.

2. We will compute all marginal effects appropriate to test the hypotheses of interest from any interaction models described below. In some cases, we may present treatment effects estimated on different subsets of the data for expositional clarity. If so, we will verify that we can reject the null of no difference in treatment effects in a more complex interactive model reported in an appendix when possible.

3. Don't know responses will be considered missing data for the factual belief outcome measures.

4. We will also compute and report summary statistics for our samples. We will also collect and may report response timing data as a proxy for respondent attention.

5. The order of hypotheses and analyses in the final manuscript may be altered for expositional clarity.