Social Debunking of Misinformation on WhatsApp: The Case for Strong and In-group Ties

IRENE V. PASQUETTO, School of Information, University of Michigan, USA and Shorenstein Center on Media, Politics and Public Policy, Harvard Kennedy School, USA
EAMAN JAHANI, Institute for Data, Systems and Society, Statistics and Data Science Center, MIT, USA
SHUBHAM ATREJA, School of Information, University of Michigan, USA
MATTHEW BAUM, John F. Kennedy School of Government, Harvard University, USA

In this paper, we argue that WhatsApp can play an important role in correcting misinformation. We show how specific WhatsApp affordances (flexibility in format and audience selection) and existing social capital (prevalence of strong ties; homophily in political groups) can be leveraged to maximize the re-sharing of debunking messages, such as those accessed by WhatsApp users via ChatBots and Tip-Lines. Debunking messages received in the format of audio files generated more interest and were more effective in correcting beliefs than text- or image-based messages. In addition, we found clear evidence that users re-share debunks at higher rates when they received them from people close to them (strong ties), from individuals who generally agree with them politically (in-group members), or when both conditions are met. We suggest that WhatsApp leverages our findings to maximize the re-share of those fact-checks that are already circulating on the platform by using the existing social capital in the network, unlocking the potential for such debunks to reach a larger audience on WhatsApp.

CCS Concepts: • **Human-centered computing** → **Social content sharing**; • **Social and professional topics** → **Computing / technology policy**.

Additional Key Words and Phrases: misinformation; disinformation; debunking; fact-checking; messaging apps; strong ties; format; WhatsApp

ACM Reference Format:
Irene V. Pasquetto, Eaman Jahani, Shubham Atreja, and Matthew Baum. 2022. Social Debunking of Misinformation on WhatsApp: The Case for Strong and In-group Ties. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 117 (April 2022), 35 pages. https://doi.org/10.1145/3512964

1 INTRODUCTION

WhatsApp is an affordable and user-friendly messaging app. The platform provides users with flexibility in terms of which format to choose for sharing content (text, audio, video, etc.), great control over the selection of their audiences, and a perceived feeling of privacy and secrecy. Also, on WhatsApp, users tend to know each other personally, suggesting a prevalence of strong ties

Authors' addresses: Irene V. Pasquetto, irenevp@umich.edu, School of Information, University of Michigan, 105 S State St, Ann Arbor, MI, USA, 48109 and Shorenstein Center on Media, Politics and Public Policy, Harvard Kennedy School, 79 JFK St., Cambridge, MA, USA, 02138; Eaman Jahani, eaman@mit.edu, Institute for Data, Systems and Society, Statistics and Data Science Center, MIT, 77 Massachusetts Avenue, Building E18-407A, Cambridge, MA, USA, 02139-4307; Shubham Atreja, satreja@umich.edu, School of Information, University of Michigan, 105 S State St, Ann Arbor, MI, USA, 48109; Matthew Baum, matthew_baum@hks.harvard.edu, John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Cambridge, MA, USA, 02138.

2573-0142/2022/4-ART117 $15.00
https://doi.org/10.1145/3512964

(with some exceptions, such as groups activated by political campaigners during election time). Taken together, these affordances have made WhatsApp the most popular messaging app and preferred means for everyday communication and information sharing in many countries, including India, where we conducted our research [87]. However, recent research suggests that WhatsApp is also an effective vehicle of mis- and disinformation, particularly in India [32, 85]. A major factor linked to the spread of misinformation on WhatsApp is its use as a political propaganda tool [28], which has been well harnessed by political parties in India [51]. One report called the 2019 Indian elections as the 'WhatsApp election' [78] noting the impact WhatsApp had on the elections, while others specifically underlined the role of misinformation on WhatsApp in deciding the outcome of 2019 elections [10, 13, 69, 83]. Given the wide penetration of WhatsApp in the Indian society, misinformation spreading through WhatsApp has had real-world consequences even beyond elections [10, 36, 90]. Amid the ongoing COVID-19 pandemic, misinformation on WhatsApp has not only caused health concerns by projecting alternative medicines as potential cure for COVID-19, but also resulted in panic situations through unverified claims about internet shutoffs and shortage of essential commodities during the lockdown period [3, 50].

Current company's interventions to counter mis- and disinformation on WhatsApp include labeling viral forwards and chain messages, limit forwards, and the design of on-platform Tip-Lines and ChatBots. Little evidence exists on whether any of these interventions work, or to what extent. Due to WhatsApp's encrypted nature, it is difficult for researchers to investigate the efficacy of any on-platform fact-checking efforts. We agree with others that it is crucial to understand what constitutes effective fact-checking before designing policies aimed at combating mis- and disinformation [26, 29, 68]. Thus, we designed an online experiment that allowed us to explore how to maximize the efficacy of current company's interventions, and in particular of fact-checking corrections (also called "debunking messages" or simply "debunks") accessed by WhatsApp users using existing Tip-Lines and ChatBots. We tested whether we can leverage WhatsApp's affordances and existing social capital to amplify the re-sharing of such debunking messages on the platform.

Specific WhatsApp affordances – the fact that the platform is primarily meant for intimate and private connections, and that users need to have a person's number to add them as a contact on WhatsApp – suggest that WhatsApp users tend to personally know their WhatsApp contacts, hence indicating the strong possibility of a prevalence of strong ties vs weak ties on the platform [48, 66]. Then, debunking messages shared via WhatsApp are more likely to come from strong ties. We tested whether debunks sent by strong ties are more likely to be re-shared, compared to debunks received by weak ties. Relevant literature suggests that, generally speaking, sources perceived as credible tend to be more persuasive than those who are not [73]. In the context of interpersonal communication, researchers argue that *goodwill* – whether a source has the receiver's best interests at heart [53, 59] – is an important dimension of source credibility. Thus, we hypothesize that strong ties – who might be perceived as credible for having the receiver's best interest at heart – would be more effective (i.e., persuasive) in convincing others to share the debunks.

The use of WhatsApp as a political propaganda tool resulted in many users being members of partisan, political groups. In this case, users might not necessarily know each other personally, but they share a political affiliation. We know that fact-checks tend to be shared in a partisan manner – to denigrate a member of a disliked party or to vindicate a member of a liked party [6, 80]. Thus, in the context of WhatsApp, political fact-checks are perhaps most likely to be received by connections who generally share the same political affiliation, but disagree about the topic of a specific fact-check. Attitude homophily has been shown to increase perceived source credibility by enhancing the trustworthiness dimension of credibility [42]. Thus, we also looked for the effect of political agreement, and for interactions between ties closeness and political agreement, as the latter might strengthen further the effect of the former.

Finally, given that WhatsApp allows users to share content in many different formats, we also tested if a particular debunking format is more likely to attract the interest of the receiver and therefore, re-shared more often. We hypothesize that richer, audio and image-based formats would be more interesting than text-based formats, and that this should result in greater belief change in cases when message contents are persuasive. To test our hypotheses, we conducted an online survey experiment in India. In addition to the widespread prevalence of misinformation on WhatsApp in India, our choice is also motivated by the fact that India is the largest market for WhatsApp [82], with WhatsApp also being the single largest social platform in India [61]. The experiment was also conducted in Pakistan – which has similar issues of mis- and disinformation on WhatsApp – for replication purposes.

2 BACKGROUND

2.1 WhatsApp's Affordances and Social Capital

WhatsApp's affordances have played an important role in establishing the popularity of WhatsApp. While enabling limited broadcast communication, WhatsApp provides users with great control and flexibility over information sharing and management overall. Even the groups on WhatsApp constitute highly regulated spaces [92] as the admins can remove or ban members at any time at their own discretion. The admins often tend to set a group's tone and, using feedback from other members, decide what is and is not allowed to be discussed within the group [92]. Therefore, when used for political campaigning, WhatsApp groups can successfully support microsegmentation and microtargeting of audiences [27].

Furthermore, contrary to social media platforms, WhatsApp is not subjected to algorithmic curation or content moderation. If users want to share a particular message, they can explicitly state and select the audience for that message. WhatsApp also provides delivery and read receipts for each message, which act as a further indicator of whether the intended audience has received/read the message. In addition to all this, users also have the ability to quickly and easily share content in different modalities (text, audio, video, images). Images are often shared in the form of "forwarded messages," which report no or little information about the original source(s) of that message or where it originated. This is unlike social media platforms such as Facebook or Twitter, where sharing or retweeting also carries information about the original post and its source - even if these metadata can be easily manipulated [2]. Past research also suggests that messaging app users tend to share and re-share information with close friends and family [48, 66]. With the exception of political campaigners, users of messaging apps typically know most of their contacts personally, as a user needs one's phone numbers to add them as their contacts - suggesting a prevalence of strong ties vs weak ties on WhatsApp.

Taken together, WhatsApp's affordances and the existing social capital have set it up as an environment in which users feel comfortable sharing information, including news that might be considered controversial (such as debunking messages) [9, 92] and that users might fear to share in other spaces. Great flexibility in terms of audience selection, and the possibility of retaining control over one's self-presentation have been found as an important motivator for sharing news online [44]. Valeriani & Vaccari [92] found that users who politically censor themselves on social media are more likely to engage in political discussions on WhatsApp. As the authors note, "network selection and message control allowed by these platforms (messaging apps) facilitate the circulation of controversial information within closed circles" [92]. Others have shown that communication on messaging apps increases as a response to distrust in mainstream media and reluctance to express oneself in online forums for fear of repercussions [9, 92]. Finally, another factor that might increase sharing of controversial information on WhatsApp might be that all exchanges on the messaging

apps are encrypted and invisible to anyone outside one's own network (including the company itself), which gives users a perceived sense of privacy and security [81].

2.2 Adoption of WhatsApp in India

WhatsApp's affordances of audience selection, format flexibility and perceived secrecy contributed to its successful adoption in many countries. In addition to these, a set of local and socio-economic factors played a major role in its widespread adoption in India, specifically. WhatsApp was one of the first messaging apps to be introduced in India in 2010 [67]. In the absence of multimedia apps like WhatsApp, people would rely on Short Message Services (SMS) to share content with others. Users would either pay for each SMS that they sent or buy bulk SMS packs from telecom providers which were both costly and came with certain limits [40]. WhatsApp provided a convenient and low-cost alternative to this – the app was free to use, and it only required a smartphone and an Internet plan. The internet prices further fell drastically, as 3G connectivity continued to increase [96]. Even for the smartphone, WhatsApp had an edge over early competitors like iMessage and Blackberry Messenger, as it supported different operating systems, including Android, which continues to be the most affordable and popular type of smartphone in India [89].

WhatsApp's popularity in India soared further with the entry of a new telecom company, Jio, which disrupted the telecom market with its low-cost mobile internet plans. According to the Mary Meeker Internet Trends Report 2017, following Jio's entry, mobile internet costs in India continued to drop by 10% per gigabyte on a quarterly basis [61]. Furthermore, Jio's feature phone, estimated to control almost half of India's feature phone market, supports WhatsApp as one of the only social apps on its operating system [82]. Consequently, WhatsApp has now become the single largest social platform in India [61], while India also continues to be the largest market for WhatsApp [82].

Other than economic factors, WhatsApp's minimalistic design has also contributed to its success in the Indian market. WhatsApp contains "no ads, games or gimmicks" [43] with simple features geared around sending and receiving messages. This makes it a particularly attractive and usable option for low-literate users or those who are not comfortable with the English language. Therefore, WhatsApp has expanded beyond a social app and turned into a one-stop place for many different needs, as small businesses are also utilising it for reaching out to customers and running their operations [79].

2.3 Spread of Mis- and Disinformation on WhatsApp

Misinformation is becoming a major problem in India. A recent report by Microsoft identified fake news as one of the major risks for Indians on the Internet. According to their report, approx. 64% of Indians were exposed to some form of misinformation on the Internet, more than any other country surveyed [63]. The problem is aggravating further as India's Internet user base continues to grow. According to the Mary Meeker report, India's internet population demonstrated a 40% year-on-year growth in 2018. Furthermore, mobile internet users in India account for a major chunk of this Internet traffic – almost 80%, while the global average is around 50% [61]. Consequently, WhatsApp – the largest mobile social platform and one of the major sources of news for Indians on the Internet [49] – has found itself at the center of the misinformation debate.

The role of WhatsApp in spreading misinformation has been a subject of many journalistic investigations as well as research studies. For instance, a study by [32] found ample evidence of misinformation spreading through WhatsApp in India. WhatsApp allows users to form groups of up to 256 members, where the group owners have the option to make their group public by sharing a web URL that can be used by anyone to join the group. According to [32], these public groups play a significant role in spreading misinformation on WhatsApp. Of particular importance is the

role of message forwards, where certain messages are mass forwarded across groups, allowing them to reach a wide audience [25].

Moving beyond technical features, the media agency BBC conducted in-depth interviews with WhatsApp users in India to understand factors that lead to ordinary people sharing misinformation on WhatsApp [85]. Their analysis focused on socio-cultural factors such as the declining trust in traditional media as well as the WhatsApp ecosystem. Since WhatsApp is primarily meant for close and intimate connections, the line between political news and other kinds of messages (everyday greetings, jokes) is often blurred and users tend to give more importance to the sender of the message relative to its content [85].

Another factor linked to the spread of misinformation on WhatsApp is its use as a political propaganda tool [28]. Given its wide reach and penetration in the Indian population, WhatsApp has been well harnessed by political parties in India. Bharatiya Janata Party (BJP), which is currently in power in India, has been particularly active in using WhatsApp for mobilising support. According to one report on the 2019 general elections in India [51], the party planned to have 3 WhatsApp groups for every polling station in India (there are over 900,000 polling stations) to be managed and run by local party volunteers for spreading campaign materials. One of the reports also called the 2019 elections as the 'WhatsApp election' [78] noting the impact WhatsApp had on the elections, while many other reports specifically underlined the role of misinformation on WhatsApp in deciding the outcome of 2019 elections [10, 13, 69, 83].

Given the wide penetration of WhatsApp in the Indian society, misinformation spreading through WhatsApp has had real-world consequences even beyond elections. In a much-documented instance from 2017 [10, 36, 90], 7 men were beaten to death by mobs who suspected these men to be child kidnappers – simply based on a WhatsApp message that warned people of out-of-state kidnappers preying on children. This is not an isolated instance, as the report by [12] identified many different cases of mob violence In India that were linked to a misinformation circulating on WhatsApp.

2.4 Spread of Mis- and Disinformation on WhatsApp Outside India

India is by no means the only country that registered issues of mis- and disinformation on WhatsApp [27, 35, 74, 76]. The most extensive research on misinformation on messaging apps has been conducted in the 2018 Brazil presidential elections. Resende et al. [74] collected over 700,000 messages circulating in public WhatsApp groups from that time, and using existing fact-checked details, identified many instances of misinformation – both textual and images. They further extracted a user network emerging from sharing patterns within WhatsApp groups that showed thousands of users connected to each other as part of a single dense cluster. The network contained certain users who were acting as bridges between two or more clusters; and a few central users who were connecting multiple groups simultaneously. The network structure showed that despite being an MIM, WhatsApp enabled a social network that is usually found on platforms such as Twitter or Facebook, thereby having the potential to make a piece of information go viral very fast.

Another factor that underlines the link between misinformation and MIMs is the increasing use of MIMs for political engagement. Different studies have noted how MIMs are being used for more than casual communication, and are becoming important venues for consuming news and other kinds of political engagement. Gill & Rojas [35] in a survey of South Korean adults found that the use of MIMs for political discussion was directly linked to higher tolerance toward misinformation. Rossini et al. [77] studied factors that contribute to the sharing of misinformation on Facebook and WhatsApp, which is denoted as dysfunctional information sharing. They identify political engagement as a significant predictor of sharing misinformation (either intentionally or unintentionally) across both WhatsApp and Facebook.

2.5 Countering Mis- and Disinformation on WhatsApp

Current interventions to reduce misinformation on WhatsApp include labeling viral forwards and chain messages, limit forwards, and the design of in-platform tip-lines and ChatBots [97]. WhatsApp introduced labels to indicate when a message is viral and that it has been forwarded to someone from another user. These labels can be applied to text, image, video and audio messages. For example, WhatsApp labels forwarded messages coming from a chain of five or more chats, meaning at least five forwards away from its original sender, with a double arrow icon [97], and enables on-platform Google searches for such forwards [98]. The latter feature raises concerns over its efficiency given that search engines have been found to be unreliable and biased sources of information [65]. WhatsApp is also experimenting with limiting the number of groups and individuals one can forward a message to. In 2019, the company began limiting the number of people you can forward a single message to five. During the COVID-19 emergency, WhatsApp announced to limit the forward of viral messages to one chat at a time [97]. In 2019, WhatsApp introduced a number of in-platform Tip-Lines and Chat-Bots managed by fact-checking organizations, NGOs and civil societies that users can use to verify information they have seen on WhatsApp about a certain topic or events. Examples include a Tip-Line for the Indian Election misinformation [45] and the Poynter's ChatBots for COVID-19 misinformation [37].

Little evidence exists on whether any of these interventions work, or to what extent. Preliminary research suggests that the forwarded flag had little impact, as most users had either not noticed it or misinterpreted what it meant [30, 85]. It was also noted that for most users, the identity of the sender mattered much more than the source of the content and the decision to share or forward the message was often based on the identity of the sender rather than the message or its source [85]. We agree with others that it is crucial to understand what constitutes effective fact-checking before designing policies aimed at combating mis- and disinformation [26, 29, 68]. Hence, building from existing WhatsApp affordances and social capital, we decided to explore what factors can maximize the circulation of fact-checking messages on WhatsApp that expose the falseness or inaccuracy of a piece of news, such as those accessed by WhatsApp users using Tip-Lines and ChatBots. Our goal is to understand how to employ WhatsApp affordances and social capital in our favor. After reviewing relevant literature (see below), we developed the following overarching hypotheses 1) certain formats might lead to re-sharing more than others; 2) strong ties might be more effective than weak ties in maximizing the re-share of debunking messages; 3) political agreement plays an important role in enabling re-share of debunks.

3 LITERATURE REVIEW AND HYPOTHESES

We reviewed relevant literature on what factors might enable or restrict online information sharing (sharing format, ties closeness, political agreement). Our research design enables us to examine whether WhatsApp users are more likely to feel comfortable re-sharing debunks if these are received:

1) In the form of text, image, or audio;
2) From a weak or a strong tie (an acquaintance or a close relationship);
3) From someone who agrees or disagrees with them politically (in-group vs out-group).

3.1 Correction Format

A swath of literature in persuasion offers suggestions as to how to make messages more convincing and shareable – and conveniently, some features can enhance both qualities. One key factor that can generate both persuasion and shareability is interest [11, 15, 41, 64, 100]. Interest drives sharing

in part due to self-enhancement motivations – sharing interesting information makes one appear interesting – and in part due to enhanced physiological arousal [15, 16, 99]. However, fact-checkers may be reluctant to focus on certain issues or claims simply because they may be more interesting to their audiences. Research conducted with U.S. fact-checkers found that fact-check coverage of issues was driven more by professional motives and journalistic values than readers' interests [38]. Young et al. [100] demonstrated that simply changing the format of a fact-check changes can enhance interest in it and thereby make it more effective. However, the research on this matter is not conclusive. Decades of research on the effects of communication modality on persuasion have pointed in opposite directions – with some researchers observing differences in text, audio and audiovisual formats [20] and others not [7, 23, 72, 84]. Others suggest that the relationship between format and persuasion may be more nuanced, depending on the credibility of the source sharing the message. Texts tend to be processed systematically, meaning that source credibility has less of an influence on message persuasiveness [18]. By contrast, source attributes tend to have more influence when a message is presented in an audio or audiovisual format – but this again is complicated. If a source is evaluated negatively, receivers judge a message on that basis alone. If a source is evaluated positively, receivers judge a message's argument quality and the source (ibid.). At the very least, however, it would seem intuitive that richer, audio and image-based formats would be more interesting than text-based formats, and that this should result in greater belief change in cases when message contents are persuasive.

Therefore, we test for the following hypotheses in our experiment to evaluate the effect of debunk message formats on their success.

- H1: Image- and voice-based corrections will generate greater message interest than text-based corrections.
- H2: Image- and voice-based corrections will change beliefs more than text-based corrections.
- H3: Image- and voice-based corrections will be more likely to be shared on social media and messaging apps than text-based corrections.

3.2 Relational Closeness of Source

Messaging apps are primarily used to connect with family and friends, as you need someone's number to add them as contacts [48, 66]. Thus, fact-checks shared via WhatsApp are more likely to come from strong ties. We hypothesize that strong ties might be effective in enabling re-sharing of debunks because they might be considered more credible than weak ties. [73] found that sources perceived as credible tend to be more persuasive than those who are not. In addition, several researchers have indicated that the credibility of contacts may be used as a heuristic for judging the credibility of content they share, particularly in cases where the receiver is overwhelmed with content or the original source information is not immediately available [47, 62]. In the context of interpersonal communication, researchers have tended to conceptualize source credibility as having two dimensions: expertise and trustworthiness (ibid.). However, other researchers have argued for a third dimension, goodwill, which refers to whether a source has the receiver's best interests at heart [53, 59]. The core idea is that strong ties – defined as individuals with whom one has a close relationship with [21] – might be perceived as more credible than weak ties because receivers might perceive them as having their best interest at heart. In addition, strong ties have been generally known to be more persuasive than weak ties in virtue of decreasing the likelihood that receivers perceive a message as intended to persuade, as persuasive intent tends to trigger counter-argument [71, 93]. As such, we test for the following hypothesis:

- H4: Corrections shared by strong ties will have larger effects on re-sharing than corrections shared by weak ties.

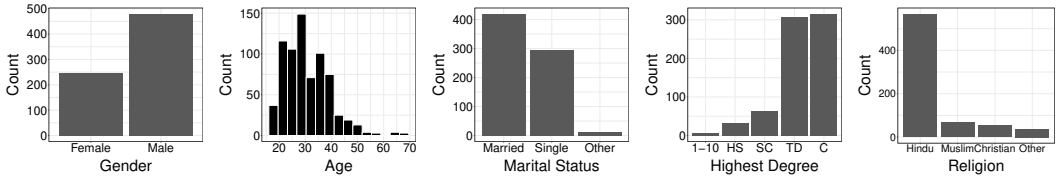


Fig. 1. The distribution of experiment participants in India across various demographic variables. In the Highest Degree subplot, 1-10 indicates less than 10 years of schooling, HS indicates High school, SC indicates Some College, TD indicates Technical Degree and C indicates College.

3.3 Attitudinal Similarity of Source and Costly Rhetoric

We know that fact-checks tend to be shared in a partisan manner – to denigrate a member of a disliked party or to vindicate a member of a liked party [6, 80]. Thus, in the context of WhatsApp political groups, fact-checks are perhaps most likely to be received by connections who generally share the same political affiliation, but disagree about the topic of a specific fact-check. Attitude homophily has been shown to increase perceived source credibility by enhancing the trustworthiness dimension of credibility [42]. When a message is relevant to an in-group identity, seeing a message from an in-group member can encourage systematic processing [54]. Source trustworthiness appears to be particularly important when correcting misinformation, even more so than expertise [39, 60]. In addition, messages generally – and corrections in particular – have been found to be more persuasive when they are offered by someone apparently speaking against their self-interests [14, 17]. For example, if a co-partisan offers a correction that shows misinformation about an opposition partisan to be wrong, that message may be perceived as more credible than if it were offered by an opposing partisan. Such "costly rhetoric" generates surprise, increases perceptions of source trustworthiness and message accuracy, and reduces argument scrutiny [71]. We thus test for the following hypothesis:

- H5: Corrections shared by an in-group member will have larger effects on re-sharing than corrections shared by an out-group member.

4 RESEARCH DESIGN

We tested our hypotheses through an interactive online survey experiment. The experiment was conducted in India, and in Pakistan for replication purposes. We first describe the recruitment process followed by the experiment flow.

4.1 Recruitment Process

All participants were residents in their respective countries and self-reported users of WhatsApp or other Messaging Apps. Research company Ipsos managed the recruitment of the participants and the administration of the surveys in close collaboration with the authors of the paper. The criteria for enrollment was twofold. First, the participant must be of the age of majority in his or her country. Second, the participant must have used at least one closed messaging application (e.g., WhatsApp, Viber, Facebook Messenger) to inform themselves about events outside their circle of friends and family – namely by consuming news – in the past week. In each country, we attempted to obtain a sample demographics mirroring that of mobile instant messenger users in that country. Recruiters visited busy places such as street markets and conducted on-street intercept screening using approved recruitment and screening documents. Potential participants provided their name, email address and phone number so they could be re-contacted by researchers. Potential participants completed a two-question, preliminary survey at the point of recruitment,

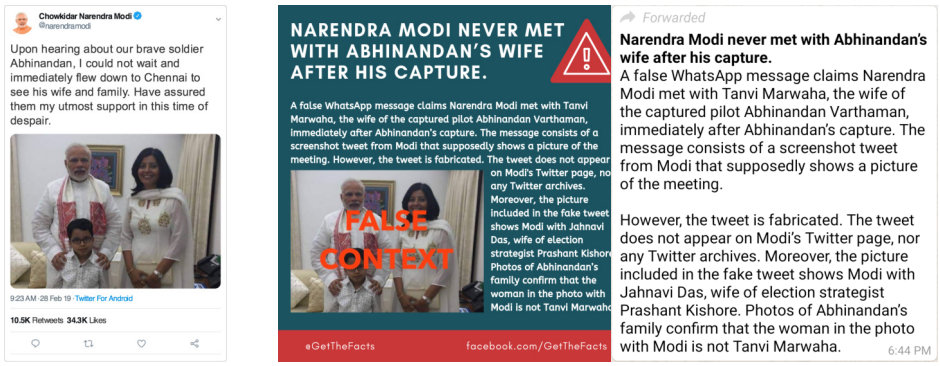
and again just prior to completing a survey or interview, to verify that they meet the criteria for eligibility. If they fit the screening criteria, recruiters furthers contacted the participants to collect their preferred electronic contact information and any relevant demographic information. A stratified random sampling of those that expressed interest in participating in the study and met the screening criteria was selected for receiving the survey. This survey was administered electronically or verbally, depending on whether the participant was being contacted online, in-person or via telephone or webcam. The experiment was approved by the local Institutional Review Board (IRB) and all participants signed an informed consent at the beginning of the experiment. After being presented with the consent form, online participants indicated consent to participate in the survey by providing an electronic signature. Participants contacted in-person, or via telephone or webcam, verbally affirmed that they consent to participating in the survey after being read the consent form. Both consent forms and surveys were administered in the local language.

Overall, we recruited and collected data from 1348 participants in India and 1457 participants in Pakistan. Figure 1 shows basic descriptive stats on demographic variables of our participants in India. Even though our sample is not representative of the larger population in each country, any estimates we find can be simply weighted and generalized to the larger population. For the remainder of this section, we describe the experiment design in India since the survey in Pakistan followed very similar steps.

4.2 Experiment Flow

Figure 3 presents an overview of how our experiment flows. We utilize a 2 (Tie Strength: Strong or Weak) by 2 (Tie Political Agreement: High or Low) by 3 (Correction Format: Text or Audio or Image) factorial design (represented by blue boxes in Figure 3) and the participants were block-randomized to one of the 12 possible conditions. First, the participants reported which of the two major political parties in India they prefer using a feeling thermometer. Then, they were asked to provide the first name of a contact that meets a two-part description, each part of which varied randomly (represented by STRENGTH-TREATMENT and GROUP-TREATMENT in Figure 3). Half of the participants were asked to name a strong tie (i.e., a relationally close individual), and the other half were asked to name a weak tie (i.e., a relationally distant individual). At the same time, half of the participants were assigned to name someone who strongly agrees with their political beliefs and the other half to name someone who strongly disagrees with their political beliefs. The random assignments into tie strength treatment were independent of random assignment into political agreement treatment. After naming a tie, all participants saw a pro-attitudinal example of misinformation in the format of a WhatsApp forward – that is, an example of misinformation they are likely inclined to believe given their stated preference toward the two major political parties in India. Figure 2a is an example of misinformation that was shown to participants who indicated preference for BJP (Prime minister Narendra Modi's party) over Indian National Congress (the major opposition party).

After viewing the misinformation, each participant was asked to imagine their tie had sent them a correction message debunking the misinformation they just saw. The debunk message was randomly assigned to be in one of three possible formats, text, audio or image. These corrections were nearly identical in terms of content. However, text-condition participants saw a simple text-message version of the correction, audio-condition participants heard someone read the correction, and image-condition condition participants saw a visually produced version of the text-based correction that also included a relevant image. Figures 2b and 2c are examples of these debunk messages in image and text format, respectively. After the participants saw the correction, they answered multiple questions designed to measure their interest in the debunk message, its credibility, and their intention to re-share it. To ensure that the type of the tie influences their decision to re-share



(a) (b) (c)

Fig. 2. The English version of misinformation and debunk messages shown to participants in favor of BJP. (a): The original misinformation piece. (b): The debunk message in image format. (c): The debunk message in text format.

the debunk message, we explicitly worded the survey question as "How likely would you be to share {name-of-the-tie}'s message on WhatsApp?".

In order to measure the efficacy of debunk messages in correcting beliefs, we asked the participants to rate their belief about the accuracy of the false claim twice, once immediately before seeing the tweet or the Facebook post containing the misinformation (e.g. figure 2a), and second after seeing the debunk message. These two points during the experiment flow are also illustrated in figure 3. The difference between these two measures the effectiveness of debunk messages in correcting the beliefs about the misinformation. Typically, experiments testing corrections introduce a distraction task of 5 to 30 minutes between the presentation of the misinformation (and the debunk) and the final belief scale (Seifert, 2002; Ecker et al., 2017). For our distraction task, we use a battery of demographic and news consumption questions that seemed unlikely to influence responses to the belief scale, and would have otherwise been asked at the end of the survey.

A major challenge with our experiment design is non-compliance to intended treatment: even if a participant is assigned to the weak tie condition and is asked to name one, they may nevertheless still mention a strong tie. A similar non-compliance can happen with the political agreement condition. Such a non-compliance could significantly reduce the power of any statistical tests. To account and correct for this non-compliance, we utilize the encouragement design in randomized experiments [8] that requires knowledge of the actual treatment, which may be different from the intended treatment. We implement this by including multiple questions at the end of the survey, asking the participants to report their level of closeness and political agreement with the tie they had mentioned at the beginning of the survey. We used these self-reported measures in our analysis as the actual treatment within the encouragement design framework. As part of these questions, tie strength was measured using the Uni-dimensional Relationship Closeness Scale (URCS), as Marsden and Campbell [58] argued that relational closeness is the best indicator of tie strength. In-group and out-group membership – in other words, degree of political agreement – was measured using the question "How much does TIE disagree or agree with your political beliefs?" (replacing TIE with the name of the tie the participants mentioned), which will be measured on a five-point scale ranging from "Strongly disagree" to "Strongly agree". The complete survey protocol can be found at this link[1].

[1]https://osf.io/4uma7/?view_only=da2e919a960a4a808ea3b82fbf5b6b3f

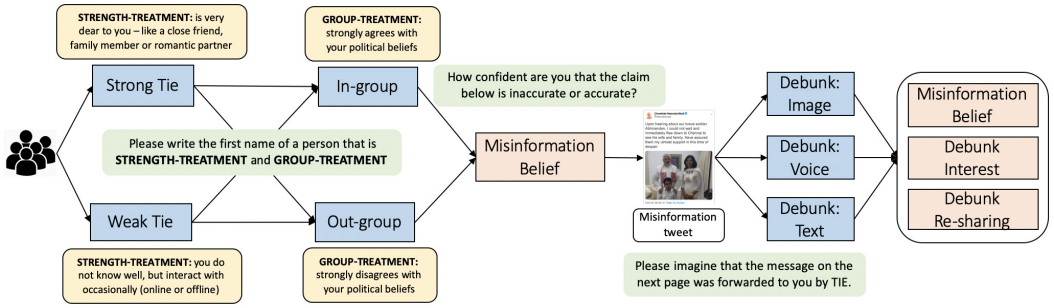


Fig. 3. The experiment design. STRENGTH-TREATMENT and GROUP-TREATMENT are variables that take on the string corresponding to the assigned treatment. TIE holds the name of the person the participant enters.

5 ANALYSIS

We now describe our approach to test the different hypotheses. Our variables of interest (i.e., the dependent variables in our regression analysis or any parametric tests) include the participants' intentions to re-share the debunk message, any changes in terms of their belief in the misinformation, and the level of interest generated by the debunk message. The sharing decision is measured as an ordered categorical variable with 5 levels (from "not at all likely" to "very likely"). The interest in debunk message is also measured as a 5-level ordered variable (from "not at all interesting" to "very interesting"). The change in misinformation belief is modeled by taking the difference in the participants' belief in misinformation before and after showing the debunk message (maximum 11 levels but can be collapsed into 3 levels).

For the hypothesis concerning correction format (H1, H2, H3), we expect our randomly assigned format to be the actual administered treatment. Therefore, the randomized setup can be used without any corrections to compare the causal effects of different debunk formats. Since our dependent variables (interest in debunk message, misinformation belief change) are ordered categorical variables, we use ordinal logistic regression [33] to model the effect of different formats. In our survey, we had also asked the respondents if they perceived the debunk message as accurate and how often they discussed political news on WhatsApp. For both questions, we recorded their responses as an ordinal variable with 5 levels. Since the misinformation involved a political message, by controlling for the overall political engagement on WhatsApp, we can obtain a better estimate for the effect of formats.

In contrast to debunk formats where the randomly selected format was the actual treatment, the assigned condition for the strength and the agreement of the tie might not necessarily match the actual treatment. As presented in the research design, we rely on the subjects to name a tie within each treatment category, which can lead to a significant degree of non-compliance. We measured this non-compliance by analysing user responses on the questions we asked (about the actual tie strength and degree of political agreement with the tie) at the end of the survey. Figure 4 illustrates how the actual treatment varies by the assigned treatment and we see that subjects often recalled and provided a strong tie with whom they agree, regardless of their treatment. Even though subjects are more likely to report a strong or an in-group tie when the randomized treatment is strong or in-group, there is nevertheless significant non-compliance in terms of subject reporting a strong or in-group tie even when the randomized assignment is a weak or out-group tie. Such levels of non-compliance would make it difficult to detect any effect our intended tie characteristics may have on re-sharing the debunk message. To address this challenge, we use the encouragement

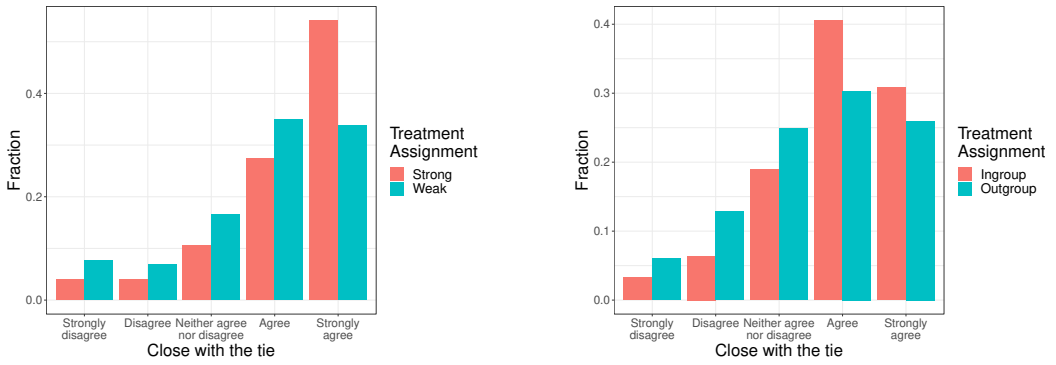


Fig. 4. The level of non-compliance with the selected treatment. (a) compares the actual strength of the tie using a self-reported measure at the end of the survey with the assigned tie strength treatment. (b) compares the actual level of political agreement with the tie using a self-reported measure at the end of the survey with the assigned tie agreement treatment.

design in randomized experiments [8, 34] which effectively treats the randomized treatment as an instrumental variable for the actual treatment. For the sake of completeness, we analyze and present the effect of tie quality in three different ways:

(1) **Self-reported measure of tie characteristics:** We compare the decision to re-share the debunk message across different levels of the self-reported tie strength or agreement. Since self-reported measures of tie strength or agreement were made using a 5-level ordinal variable, we use ordinal logistic regression to evaluate the significance of any effect. These results will not have a causal interpretation as they are not based on a randomized assignment but nevertheless would be suggestive of any potential effect.
(2) **Intention to Treat Design:** We use the randomized treatment assignment to find the causal effect of tie strength or agreement. While the findings would have a causal interpretation, they would only measure the effect of our intention to treat rather than the actual treatment, owing to the non-compliance highlighted above. As such, they would underestimate the size of the actual effect. Similar to the above analysis, we use ordinal logistic regression for the estimation.
(3) **Encouragement Design:** We use the self-reported measures of tie strength or tie agreement to infer the actual treatment on tie characteristics. If the subject agrees or strongly agrees with the statement "My relationship with the {name-of-the-tie} is strong", we label that as an actual strong tie treatment, otherwise as weak. Similarly, if the subject answers the question "How much does {name-of-the-tie} disagree or agree with your political beliefs?" with "agrees" or "strongly agrees", we label that as an in-group tie, and out-group otherwise. Given these inferred actual treatments, we estimate the actual causal effect of the tie treatment using the randomized treatment as an instrument for the actual treatment. For simplicity (also due to lack of any R package that supports logistic regression with instrumental variables), we converted the 5-scale sharing decision to a numeric integer variable ranging from 0 to 4 and estimated the model with this numeric variable using the regular linear 2 stage least squares.

We control for perceived accuracy of the debunk message and political engagement on WhatsApp in all models above to reduce the standard error in our estimates. In all of our analysis, we further

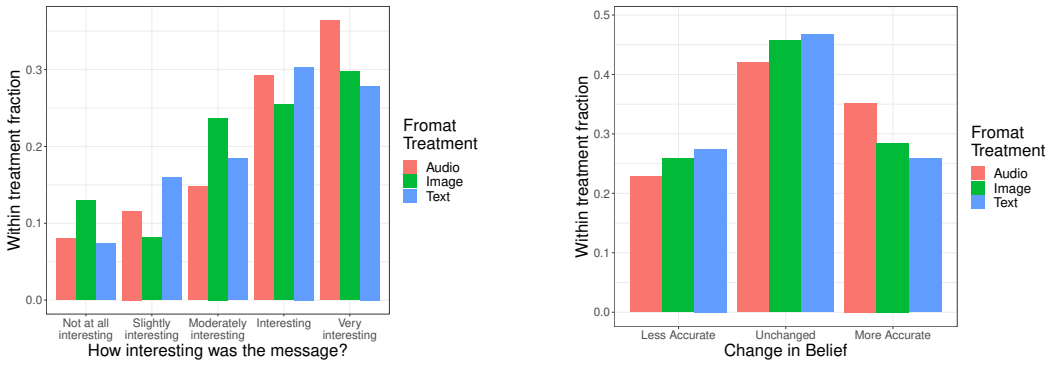


Fig. 5. (a) compares how interesting the participants found the debunk message across different formats. (b) compares the level of belief change on the misinformation after observing the debunk message across different formats. For simplicity, the levels of belief change are collapsed into 3 categories.

restricted our data to those participants who answered at least one attention check question correctly. Almost half of our participants failed to pass any of the three attention check questions we had inserted at different points of the survey, and thus our final data size was restricted to $N = 726$.

5.1 Follow-up Interviews

We conducted follow-up, semi-structured interviews and chat texting with 25 survey participants who volunteered to be contacted after the survey. Nearly 100 participants expressed an interest in participating in the follow-up process. Of those, 15 individuals agreed to be formally interviewed, and 10 to informally chat via WhatsApp with the researchers. The overall number of study participants for this follow-up phase of the study is 25, of which 9 identified as women and 16 as men. The 25 participants varied in age from 20 to 70 years old with most participants between 25 and 50 years old. None of the participants was compensated for their participation in the follow-up interviews or chat and all participants agreed to talk and chat with the researchers in English. The semi-structured interviews were conducted through WhatsApp video calls. With those who agreed to talk to us via chat, we engaged via WhatsApp one-to-one chat on a daily or weekly basis for about two weeks each. We did not ask group administrators to add us to their groups, but we did ask them to describe their groups and forward relevant content shared in them. After official interviews ended, some participants kept sending us examples of misinformation and harmful speech as they encountered these on WhatsApp for about six months. These informal interactions with the participants were documented via field memos and notes, and the messages were archived mostly via screenshots. In the discussion section we will report on selected findings from this qualitative analysis in order to critically engage and better interpret our quantitative findings.

Table 1. Ordinal Logistic Regression using Randomized format treatment. Columns (1) and (2) use the interest in the debunk message with 5-levels as the ordinal dependent variable. Columns (3) and (4) use the change in belief about the misinfo with 11-levels as the ordinal dependent variable. Columns (2) and (4) include the categorical frequency of news discussion on WhatsApp as a control variable. The intercept corresponds to "Never Discuss News"

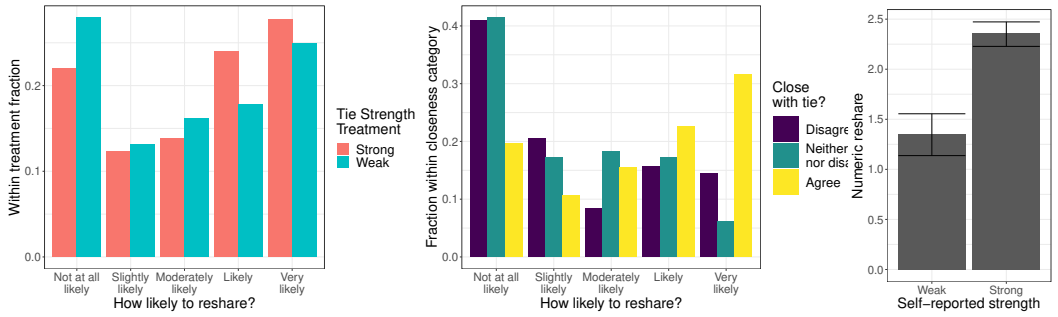
	Dependent variable:			
	Interest in Debunk Message		Misinfo Belief Change	
	(1)	(2)	(3)	(4)
Image Format	-0.340^{**}	-0.349^{**}	-0.336^{**}	-0.344^{**}
	p = 0.039	p = 0.041	p = 0.046	p = 0.041
Text Format	-0.311^{*}	-0.361^{**}	-0.446^{***}	-0.475^{***}
	p = 0.055	p = 0.032	p = 0.008	p = 0.005
Rarely Discuss News		1.095^{***}		-0.538^{*}
		p = 0.0001		p = 0.059
Sometimes Discuss News		1.516^{***}		-0.523^{**}
		p = 0.000		p = 0.047
Frequently Discuss News		2.397^{***}		-0.281
		p = 0.000		p = 0.290
Very frequently Discuss News		3.837^{***}		-0.677^{**}
		p = 0.000		p = 0.013
Observations	726	726	726	726
Akaike Inf. Crit.	2,201.905	1,980.174	2,447.055	2,446.615
Note:				*p<0.1; **p<0.05; ***p<0.01

6 FINDINGS

6.1 H1 & H2 & H3: Audio corrections generated more interest and higher belief change than text or image-based corrections. However, we could not conclude that these differences translate to more re-sharing.

Figure 5a compares the levels of interests generated by each correction format. We observe that audio formats create more interest than image or text. The chi-squared indicates that the level of interest and the (three-level) correction format are not independent ($p = 0.01$). Performing the same chi-squared test pairwise, we find that the differences between text and image formats, and text and audio formats are statistically significant ($p = 0.015$ and $p = 0.02$ respectively), but we don't observe significant differences between audio and image formats ($p = 0.22$).

Similarly, we found that audio corrections are more effective than text- or image-based corrections in changing beliefs about the misinformation claim. Figure 5b compares the effectiveness of each format on correcting beliefs. For simplicity, we have collapsed the 11-level ordinal variable corresponding to belief change, ranging from 5 levels toward less accuracy to 5 levels toward more accuracy, into three levels. The chi-squared test of independence using the original scale of belief change suggests that format and belief change are dependent with marginal significance ($p = 0.09$). The pairwise chi-squared tests indicates that only the difference between text and audio formats are statistically significant ($p = 0.039$). The chi-squared does not indicate significant differences between image and text and image and audio formats.



(a) (b) (c)

Fig. 6. (a) compares decision to reshare the debunk message over tie strength treatments, with post-stratification using frequency of news discussion on WhatsApp. (b) compares decision to reshare the debunk messages over different levels of self-reported tie strength. The legend key is the self-reported tie strength and it corresponds to how much the participant agrees with the statement: "I am close with the tie". "Strongly (dis)agree" and "(Dis)agree" categories are collapsed into a single level. (c) compares the mean decision to reshare in numeric form (ranging from 1 to 5) between different levels of self-reported tie strength collapsed into 2 levels. Bars correspond to 95% confidence interval.

Finally, we did not observe that increased interest or more accurate belief translates to higher likelihood of re-sharing the debunk. While we do see that the participants are more likely to re-share audio formats with their friends than image or text, the differences were not statistically significant. Table 1 shows the same results we iterated above, but using the ordinal logistic regression model which in contrast to the chi-squared test takes the ordering of interest levels and belief change into account. The results from ordinal logistic regression confirm the same findings as above: audio formats cause more interest and more accurate beliefs about the misinformation than text or image. The coefficients suggest that image formats are more successful than text, but the differences are not large enough to be significant.

6.2 H4: Corrections received by a close friend or family member are re-shared more than corrections received by a casual acquaintance

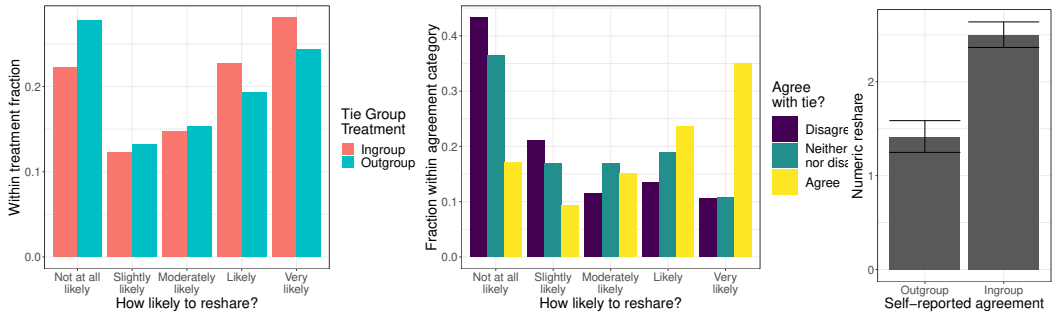
Figure 6 compares the extent of re-sharing (the debunk message) across various measures of tie strength. Sub-figure 6a does this by comparing the post-stratified frequency of re-sharing over the randomized tie-strength assignment. The post-stratification is performed using the categorical frequency of news discussion on WhatsApp, the same variable also used in Table 1. We use the post-stratification method since the frequency of news discussion accounts for a large fraction of variance in re-sharing and including it in any model makes it easier to detect the effect of tie strength. Cochran–Mantel–Haenszel test with news discussion as the control category suggests randomized tie strength treatment and re-sharing are dependent with marginal significance ($p = 0.078$). Collapsing the re-share from 5 into 2 categories leads to a much stronger result ($p = 0.007$). Sub-figure 6b compares re-sharing across different levels of self-reported tie strength after collapsing its 5 categories into 3 for better visualization (Chi-squared test $p < 10^{-10}$ for both original and collapsed measures of self-reported tie strength). In sub-figure 6c, we convert the categorical re-sharing variable to an integer variable ranging from 0 (Not at all likely to re-share) to 4 (Very likely to re-share) and compare its mean across the binary self-reported tie strength, as explained in section 5 (t-test $p < 10^{-10}$). We observe that the effect of tie strength on re-sharing is stronger and more visible when we use self-reported measures of tie strength. This points to the issue of

Table 2. The effect of tie strength on intention to re-share the debunk message. Columns (1) and (2) correspond to the intention to treat causal effect, which uses the randomized tie strength assignment in an ordinal logistic regression model. Columns (3) and (4) correspond to the causal effect of tie strength using the encouragement design, where the randomized tie strength assignment is used as an instrument for the self-reported tie strength in a linear model. Note that coefficients in columns (1) and (2) are not comparable with coefficient in columns (3) and (4) since one uses logistic regression and another uses linear regression.

	Dependent variable: Re-sharing the debunk			
	Intention to Treat Model		Encouragement Design Model	
	(1)	(2)	(3)	(4)
Weak Tie	-0.405^{***} p = 0.005	-0.493^{***} p = 0.001	-1.760^{**} p = 0.011	-1.807^{***} p = 0.008
Slightly Accurate Debunk	1.169^{***} p = 0.00001	1.175^{***} p = 0.00002	0.339^{**} p = 0.050	0.280^{*} p = 0.095
Moderately Accurate Debunk	2.221^{***} p = 0.000	2.104^{***} p = 0.000	0.946^{***} p = 0.00000	0.835^{***} p = 0.00001
Accurate Debunk	3.438^{***} p = 0.000	3.194^{***} p = 0.000	1.794^{***} p = 0.000	1.588^{***} p = 0.000
Very Accurate Debunk	5.358^{***} p = 0.000	4.772^{***} p = 0.000	2.411^{***} p = 0.000	2.039^{***} p = 0.000
Rarely Discuss News		0.586^{*} p = 0.081		0.211 p = 0.271
Sometimes Discuss News		1.182^{***} p = 0.0002		0.514^{***} p = 0.007
Frequently Discuss News		1.844^{***} p = 0.00000		0.749^{***} p = 0.001
Very frequently Discuss News		2.336^{***} p = 0.000		0.919^{***} p = 0.0002
Constant			1.323^{***} p = 0.00002	0.934^{**} p = 0.011
Observations	726	726	726	726
Akaike Inf. Crit.	1,822.684	1,752.746		
Note:				*p<0.1; **p<0.05; ***p<0.01

non-compliance as we discussed before and the difficulty in extracting the causal effect with our randomized instrument even though the effect might be large. Nevertheless, both figures offer clear evidence that participants would rather re-share corrections when they receive them from a person close to them, such as a family member or a close friend (i.e., strong tie), than from a casual acquaintance (i.e., weak tie).

Table 2 presents these results in terms of regression coefficients. We estimate the effect of tie strength using two models. The first model (columns 1 and 2) estimates the intention to treat effect where we fit the categorical decision to re-share with an ordinal logistic regression model. The



(a) (b) (c)

Fig. 7. (a) compares decision to reshare the debunk message over tie group treatments, with post-stratification using frequency of news discussion on WhatsApp. (b) compares decision to reshare the debunk messages over different levels of self-reported tie group. The legend key is the self-reported tie group and it corresponds to the participants answer to the question: "How much does the tie disagree or agree with your political beliefs?". "Strongly (dis)agree" and "(Dis)agree" categories are collapsed into a single level. (c) compares the mean decision to reshare in numeric form (ranging from 1 to 5) between different levels of self-reported tie group collapsed into 2 levels. Bars correspond to 95% confidence interval.

second model (columns 3 and 4) uses the randomized tie strength treatment as an instrument for the actual tie strength, where we fit the integer re-sharing variable (ranging from 0 to 4) with a linear instrumental variable model. Models 1 and 3 control for the perceived credibility of the debunk message and models 2 and 4 also control for the frequency of news discussion on WhatsApp. We need to control for these covariates since the standard errors are too large in the most basic model with only tie strength. Overall, the results in Table 2 confirm our hypothesis that debunk messages received from strong ties are more likely to re-shared. In contrast to models (1) and (2), the coefficient values in columns (3) and (4) are easier to interpret and they suggest that the effect size of tie strength is comparable to that of credibility of debunk or engagement in political discussion.

6.3 H5: Corrections received by a like-minded individual are re-shared more than corrections received by an unsympathetic individual

We present the results in this section in a manner similar to the previous section. Figure 7 compares the extent of re-sharing (the debunk message) across various measures of tie agreement. Sub-figure 7a does this by comparing the post-stratified frequency of re-sharing over the randomized tie-group assignment. The post-stratification is performed using the categorical frequency of news discussion on WhatsApp. Cochran–Mantel–Haenszel test suggests significant dependence between random tie-group treatment and re-sharing ($p = 0.037$ with binary re-sharing and $p = 0.22$ with the original 5 levels). Sub-figure 7b shows this comparison with self-reported tie agreement when collapsed into 3 categories (Chi-squared test $p < 10^{-15}$ for both original and collapsed measures of self-reported tie agreement). Sub-figure 7c converts the decision to re-share into an integer variable ranging from 0 to 4 and compares its mean across the binary form of tie agreement as explained in section 5 (t-test $p < 10^{-15}$). Similar to tie strength, the effect of tie agreement on re-sharing is stronger and more visible when we use self-reported measures of tie group (suggesting a great deal of non-compliance on tie-agreement treatment). Nevertheless, both figures offer clear evidence that participants would rather re-share corrections when they receive them from a like-minded

Table 3. The effect of tie group on intention to re-share the debunk message. Columns (1) and (2) correspond to the intention to treat causal effect, which uses the randomized tie group assignment in an ordinal logistic regression model. Columns (3) and (4) correspond to the causal effect of tie group using the encouragement design, where the randomized tie group assignment is used as an instrument for the self-reported tie group in a linear model. Note that coefficients in columns (1) and (2) are not comparable with coefficient in columns (3) and (4) since one uses logistic regression and another uses linear regression.

	Dependent variable: Re-sharing the debunk			
	Intention to Treat Model		Encouragement Design Model	
	(1)	(2)	(3)	(4)
Outgroup Tie	-0.308^{**} p = 0.031	-0.354^{**} p = 0.015	-1.456^{**} p = 0.015	-1.466^{**} p = 0.014
Slightly Accurate Debunk	1.111^{***} p = 0.00003	1.095^{***} p = 0.0001	0.217 p = 0.248	0.167 p = 0.352
Moderately Accurate Debunk	2.183^{***} p = 0.000	2.044^{***} p = 0.000	0.981^{***} p = 0.00000	0.889^{***} p = 0.00000
Accurate Debunk	3.413^{***} p = 0.000	3.154^{***} p = 0.000	1.754^{***} p = 0.000	1.571^{***} p = 0.000
Very Accurate Debunk	5.284^{***} p = 0.000	4.676^{***} p = 0.000	2.296^{***} p = 0.000	1.941^{***} p = 0.000
Rarely Discuss News		0.563^{*} p = 0.092		0.169 p = 0.359
Sometimes Discuss News		1.190^{***} p = 0.0002		0.389^{*} p = 0.060
Frequently Discuss News		1.809^{***} p = 0.00000		0.641^{**} p = 0.011
Very frequently Discuss News		2.307^{***} p = 0.000		0.857^{***} p = 0.001
Constant			1.453^{***} p = 0.0001	1.124^{**} p = 0.014
Observations	726	726	726	726
Akaike Inf. Crit.	1,826.026	1,758.111		
Note:				*p<0.1; **p<0.05; ***p<0.01

individual, a person they agree with politically (i.e., in-group), rather than from someone who has opposite views (i.e., out-group).

Table 3 presents these results in terms of regression coefficients. We use the exact same models and methods as in table 2. The first model (columns 1 and 2) estimates the intention to treat effect using an ordinal logistic regression model. Columns 3 and 4 use the encouragement design with a linear instrumental variable model on the re-sharing variable converted to an integer scale. Similar to tie strength, we need to control for the frequency of news discussion and debunk credibility, since the standard errors are too large in the most basic model with only tie agreement. Overall,

the results in Table 3 confirm our hypothesis that debunk messages received from in-group ties are more likely to be re-shared than out-group ties.

7 DISCUSSION

In what follows, we discuss the specifics of the effects and interactions that we found and we point at possible interventions that can be implemented starting from our results. We then present the research design limitations, reflect on issues related to the generalizability of our findings, and propose possible directions for future work.

7.1 Re-sharing Debunks in Audio Format

Debunking messages received in the format of audio files generated more interest than text- or image-based messages, and they were more effective in correcting beliefs about the misinformation stimulus. These findings add evidence to the theory that changing the format of a fact-check can enhance interest in it and thereby, make it more effective [100]. Multiple rationals might explain the finding. One of these could be users' familiarity with audio format (participants are habitual users of WhatsApp or other messaging apps). WhatsApp is one of the few platforms to allow users to send voice messages. Users can record and listen to voice messages while doing other activities. WhatsApp allows users to lock the recording bottom until the recording is over, freeing hands for other activities. Users can record multiple chains of voice messages per day, which can be auto-played in sequence, without interruption. Therefore, habitual WhatsApp users might perceive audio messages as an effortless way of consuming information, one they are familiar with outside of the experimental setting [22, 57]. High interest in audio format might also be driven by a more direct human interaction with audio than text or image. It might also be that voice-based corrections give a more personal experience than text or image based corrections that do not directly involve another individual. Then, there seems to be an opportunity for audio files to be used as effective means for fact-checking misinformation on WhatsApp. For example, institutional ChatBots and Tip-Lines that operate on WhatsApp could produce and circulate audio-based debunking messages that users can easily access whenever they are exposed to misinformation on the platform.

More research is nevertheless needed to clarify this finding. For both interest and belief change, only the difference between text and audio formats is statistically significant. In addition, the tendency of re-sharing audio formats more than other formats was also not significant. These findings might be explained by many factors, such as differences across sub-populations. Information that is easier to process is generally perceived as more familiar, and therefore more valid [5, 70]. It might be that audio files are perceived easier to process than images by certain sub-populations, but not by others. It might also be that differences exist between sub-populations in terms of propensity to believe and share fact-checks. For example, older age is associated with higher sharing of fact-checks [6], which opens up opportunities for amplifying debunking messages with the help of such sub-population. Because of visual impairments, it might be that the older adults prefer to be exposed to and re-share fact-checks in audio formats rather than in text or images. Further studies could focus on whether and why different populations might prefer different formats. Overall, in-depth research on audio formats and their role in spreading either misinformation or debunking messages on messages apps, and in particular on WhatsApp, is in its infancy [57].

7.2 Re-sharing Debunks Received by Family Members and Close Friends

We also found that our participants preferred to re-share corrections when they received these from a person close to them, such as a family member or a close friend (i.e., strong tie), than from a casual acquaintance (i.e., weak tie). This time, our hypothesis was confirmed. There might be multiple explanations for this finding. The credibility of contacts is often used as a heuristic for

judging the credibility of content they share [47, 62]. We know that strong ties have been shown to be more persuasive than weak ties because they decrease the likelihood that receivers perceive a message as intended to persuade, which tends to trigger counter-argument [71]. "Goodwill," which refers to whether a source has the receiver's best interests at heart, might enhance the credibility of a source of information [53, 59]. Thus, a person close to the participant might be perceived as more credible than an acquaintance - if participants assume that this person has the other person's best interest at heart. Generally speaking, individuals are more receptive to corrections from members of their social circles [94] or from people that they already know [31, 56].

The large presence of strong ties on WhatsApp was confirmed in our follow-up interviews. All interviewees confirmed to mainly use WhatsApp to communicate with family and close friends, via either one-to-one interactions or in small-group chats. Below, we report a few quotes from the interviews that well-represent this finding:

> *"I'm in four or five groups total on WhatsApp. One is my family group, the immediate family, another one is for high school friends (we formed the group two years ago, 46 years after graduation), one is for my wife's family, one for university friends (the 1974 graduating class)."* (Study participant n.1, February 2020)

> *"I use WhatsApp to communicate with my family members and friends. We have one local group for the family, mainly used for sharing information. I ask about their difficulties when I'm not at home. I travel quite a lot. [...] I'm in three groups, all about family or close friends... One is family, one for my community, relatives, other extended family, cousins."* (Study participant n.2, February 2020)

> *"WhatsApp is a good tool for communication because it is a free app that every friend of mine has in their hand at any given time. I receive messages only from people I am in touch with and not everyone who wants to contact me or send anonymous messages. Again, it is so easy to voice call, video call, and even voicemail or send and receive photos/pictures whenever [...] Compared to other apps, WhatsApp is easy, quick, reliable, and can be used by anyone at any time to communicate with people you know."* (Study participant n.10, February 2020)

Interviewees were generally enthusiastic about WhatsApp and reported feeling gratitude: it enables them to talk with their loved ones at no cost, which is perceived to be a great advantage, especially by older generations.

> *"I like WhatsApp very much. I like that I do not have to pay anything to talk with people, only my Internet bill. My wife can call my daughter in Europe every day; that would have been very expensive years ago. Life was terrible those days. [...] I have seen my grandson growing up on WhatsApp. It is a very useful tool. I'm glad it exists; I use it at least two or three hours per day."* (Study participant n.1, February 2020)

Overall, this might be a precious finding for fact-checking operations. Given that information shared via WhatsApp is more likely to come from strong ties, then WhatsApp can constitute the ideal space for planning convincing fact-checking interventions. The main challenge would be to find ways to track the effective virality of fact-checks on WhatsApp and how these might move from a user's closed network of contact to another. Collaboration with the platform itself might be necessary to achieve this research goal.

7.3 Re-sharing Debunks Received by Like-minded Individuals

Similarly, we found that participants would rather re-share corrections when they receive them from a like-minded individual (i.e., in-group), rather than from someone who has opposite views (i.e., out-group). This finding adds to previous work showing that uses are more likely to view a source as trustworthy if they share similar traits [86]. Thus, political groups on WhatsApp – generally speaking, environments in which users agree with each other politically, but they do not necessarily personally know each other – might also be fertile ground for effective debunking or even pre-bunking operations. For example, if a member of these groups offers a correction that shows misinformation to be wrong, that message may be perceived as more credible than if it were offered by someone external to the group. Such mechanism might work even better in case of a group member sharing a fact-check about an individual with the same political affiliation. Past studies found that corrections become more persuasive when they are offered by someone apparently speaking against their self-interests [14, 17]. This means that, potentially, fact-checks shared among members of the same political group (e.g., individuals who generally agree with each other politically, but disagree on a specific topic of a fact-check), have great chances to be taken seriously and become viral.

7.4 The Role of Source Closeness and Similarity in Debunks Re-sharing Behavior

Finally, we found that when both conditions are met – users received a debunking message from a strong ties who also agrees with them politically – they are more likely to re-share such corrections than when only one condition is met (difference is statistically significant). In other words, the strongest effects appear when the user who shares the debunk is both a close and an in-group tie. Taken together, our results suggest that perceived closeness and similarity to the source of the fact-check, either in terms of personal relationship or general political agreement or both, can result in higher chances that users re-share debunking messages. This is in line with previous research on the "virtual bystander effect," which suggests that users are less likely to fact-check in social settings in which they perceive to have no or little control over audience selection [46]. Because of its affordances of flexible audience and format selection, perceived secrecy and existing social capital in which users mostly communicate with strong ties, WhatsApp could be a unique environment for effective fact-checking operations. Our results strongly suggest that good debunking practices start with those users we already share a connection with. As discussed, fact-checking agencies could leverage these findings in multiple ways, first of all by encouraging ChatBots and Tip-Line users to re-share debunks with close friends and family on WhatsApp, or in political groups. However, we need to consider the possibility of selective exposure [1]. Any fact-checking operation that relies on users-driven rather than algorithmic-driven corrections might result in such bias as users cherry pick favorable fact-checking messages to share with others based on their own interests, and, consequently, those whose received them may be exposed to only a narrow, unrepresentative set of available corrections.

Another significant challenge emerged in the follow-up process. Interviewees noted that while they are willing to fact-check WhatsApp forwards for their own sake, they rarely engage in correcting other users by sharing evidence. When asked if they correct others on a daily basis, most users expressed the desire to do so. However, when asked to provide a concrete example of such behavior, only two respondents were able to provide it. This specific finding might help with the interpretation of results from survey research based on self-reported correction behavior that found that corrections are quite common on WhatsApp, suggesting that such pro-social behavior might be over-reported [76]. When asked to elaborate further on the challenges of correcting others, participants indicated that correcting others might be considered impolite or rude because

of cultural factors, especially if the sharer is senior or "outranks" them in terms of social status. One participant even suggested that the onus should not be on users to address information problems with the app but on WhatsApp itself.

> *" I correct only people I personally know; otherwise, I do not. People are lazy; they just forward whatever they think they agree on. They don't care if it is real or not. [. . .] I think it is WhatsApp's responsibility to clean up these messages, not mine. Otherwise, their [WhatsApp's] credibility will go down; people are getting tired of WhatsApp. There is no mechanism for fact-checking; it is not right."* (Study participant n.5, February 2020)

Both WhatsApp and fact-checking agencies should take this finding into account when developing appropriate interventions. More research is needed to understand how cultural factors such as social sanctioning impact fact-checking practices.

7.5 Comparison With Studies From Other Regions

The problem of misinformation on MIMs, especially WhatsApp, is not unique to India but has been studied in other regions as well [27, 35, 74]. Many of these studies further allude to the closed and intimate nature of MIMs (particularly WhatsApp) as a distinct feature that determines the spread of misinformation. For instance, Evangelista et al. noted how the closed and trust-based relations on WhatsApp contributed to the spread of misinformation during the 2018 elections in Brazil. Similarly in South Korea, the use of MIMs for political discussion and a high degree of homophily in one's MIM network was linked to higher tolerance toward misinformation [35]. Another study in Chile further found how WhatsApp created these mixed-communication channels where sharing of personal information was often intermixed with the sharing of political information between close ties [91].

In many ways, our findings rely on some of the similar affordances to increase the effectiveness of correction messages as well. In fact, some of it may be happening already. For instance, another study in Brazil also found that WhatsApp users were more likely than Facebook users to perform, experience, and witness social corrections to misinformation [76]. The findings were again linked to the various affordances of WhatsApp, such as the selective control over the audience of a message as well as its inherently private nature. In a different case, WeChat in China supports fact-checking services within the app itself. However, a study found that even after 2 years of adding the feature, many users were unaware about it [52]. Even those who were aware viewed fact-checked news as detached from their personal life. The authors noted that when the news was misaligned with the users' personal interests, the veracity seemed less important. Our findings help put these results in context by specifically exploring how the sharing of corrections via MIMs could become effective.

7.6 Limitations

Overall, we find our results relatively convincing, for two main reasons. First, for the most part, our findings confirm results from past studies on the nature of fact-checking interventions. Second, our design presents good indications of internal validity, as we were able to replicate the findings in another country, Pakistan (see appendix). However, the study presents some methodological limitations.

(1) **Non-compliance:** As mentioned earlier, we observed that there was a skew toward naming strong or in-group ties even if a subject was randomly assigned to the weak or out-group treatments. This observation makes sense as participants immediately think of their close contacts no matter the qualifications we asked for. This issue however has been addressed within the context of randomized experiments as the encouragement design. This design requires observing the actual treatment, but it is very challenging for us to tell whether the

named tie is actually strong or not. We attempt to infer the actual tie strength or group by asking an array of questions on strength and political agreement at the end of the survey. However, this approach is not accurate for two reasons: the definition of tie strength could be subjective among our participants and even if we could correctly deduce the strength or the group of the tie, the self-reported measures could still be inaccurate. While being a major limitation in how we measure tie strength, we believe this issue should not change our findings or at least its direction and if anything under-estimate the effect of tie strength and agreement. This is mainly because no matter what approach we used to label self-reports as strong/weak or in-group/out-group, our results based on self-reports showed very strong effect sizes as shown in the right panel of figures 6 and 7. Future studies can address these shortcomings by relying on actual observed values on the frequency of interactions (for tie-strength) and the ideological conformity of shared content (for tie-group) on social media.

(2) **Subject Heterogeneity:** Our treatments were political in nature and our subjects arguably vary greatly in terms of interest in politics. This could greatly limit their engagement with our experiment and introduce measurement errors. Lack of interest or engagement, in particular in our case with surveys, is a challenging problem which is often dealt with by collecting large samples. Given a large sample, one could either discard subjects who were not engaged or ideally estimate the effect size heterogeneity by interest or attention. We tried to address this problem by first discarding subjects who failed obvious attention check questions as they constituted a large portion of our sample and their inclusion in the sample would have nullified any effect since they were most likely providing random answers. This approach would have been similar to a model with the level of attention as an interaction term. Second, we tried to evaluate the effect of interest on the remaining subjects by testing a model with interaction terms. We found that interest plays a big role in resharing, however we did not find any significant effect on the interaction terms potentially due to our small sample size. Moreover, we found that the effect of interest is mediated through the frequency of news discussion as a model with both terms nullifies the effect of interest. Another potential issue with our analysis is that we only measured interest on the debunk message and treated it as a proxy for interest on the misinformation itself and its topic. Thus we might be missing potential heterogeneities by the topic. Overall, we were not able to fully capture potential heterogeneities among subjects by their interest level most likely due to our small sample and measurement errors. Our results should only be interpreted in terms of average treatment effects, in presence of underlying heterogeneity.

Also, in terms of study's participants, our sample tends to be skewed towards college educated individuals self-identified as binary, male and Hindu practitioners. Our study might also present limitations in terms of external validity. Ideally, the experiment should be replicated on-platform, which is currently not an option for researchers. Like others, we remain hopeful that - going forward - platforms will demonstrate greater willingness to collaborate and exchange data with misinformation researchers [4]. We consider our understanding of how and when debunking operations might be effective on WhatsApp is in its preliminary stages.

7.7 Generalizability

WhatsApp and Messenger's user base has quadrupled globally since 2014 [95]. Several reports released by marketing agencies show that messaging apps' user base grew further during the pandemic [88]. In March 2020, Facebook reported that total messaging in those countries hit hardest by the virus increased by more than 50 percent that month [88]. Also over the last year or so, deplatforming and shadowing operations are becoming common means to curb misinformation

on social media, a factor that seems to have encouraged deplatformed users and their audiences to migrate to unmonitored spaces like encrypted messaging apps [75]. Signal saw a 677 percent increase in downloads in January 2021 and Telegram's downloads increased by 146 percent over the same period of time [55]. Thus, understandably, the rise of messaging apps is causing growing anxieties among fact-checkers and media stakeholders, who struggle to find effective ways to identify and debunk false information on these spaces [55].

In response to these concerns, we suggest that fact-checking interventions should be tailored on messaging apps' user bases, affordances, contexts and regions of adoption. Our findings offer some insights on how to maximize the impact of fact-checking operations on WhatsApp: by exploiting the large presence of credible and familiar information sources on WhatsApp (strong ties, in-group members) to increase the re-share of debunking messages as these are originally shared by institutional ChatBots and Tip-Lines. The generalizability of our findings to other messaging apps and fact-checking solutions will depend on a series of factors. Marketing reports and anecdotal observations strongly suggest that different messaging apps have different user bases, are typically employed for different information practices, and that both user bases and sharing practices vary by region. WhatsApp is predominantly used in Europe and the Global South for communicating with family and friends, or for political campaigning during election time [87, 88, 95]. Facebook Messenger is also mostly used to communicate with strong ties in the US and in some regions in Asia, such as Myanmar and the Philippines. Thus, our findings might be informative also for debunking operations implemented on Facebook Messenger. Telegram and Signal, however, seem to be the preferred tools of communications for advocacy and political organizing, but not for one-to-one interactions with close family members and friends [24, 75]. Thus, Telegram and Signal might not have the same large presence of strong ties that WhatsApp and Messenger have. Still, our findings on the effect of in-group members on re-sharing behavior might apply also in the case of Telegram and Signal, specifically within small groups of like-minded individuals.

In addition, while WhatsApp and Messenger limit the number of members for public groups to a few hundreds, Telegram features public channels that offer broadcast functionalities and enable channel admins to reach up to 100.000 users. Then, Telegram can be better understood as a hybrid-platform that exists at the intersection of social media and messaging apps. As observed by Roger [75], Telegram, because of its combination of private chats and public channels, has both the reputation and the right affordances that are attractive to those seeking "to retain control over what is known about oneself while still participating (and becoming popular) on social media," a concept also referred to as "social privacy" [19]. Future experiments could focus on how to employ the affordance of social privacy to combat misinformation specifically on Telegram.

Finally, also fact-checking interventions using ChatBots and Tip-Lines vary by platforms and regions. For example, Facebook has implemented fact-checking partnerships with local agencies in most countries where either WhatsApp or Messenger are popular. These include India but also Canada, Australia, New Zealand, Iraq, Libya, Egypt, Norway, and Greece [88]. But in other countries like Afghanistan, Nigeria, Pakistan and South Sudan, they do not have partners. Telegram has not implemented any partnership with fact-checking agencies yet (at least publicly).

Messaging apps are not all the same, and different affordances might result in different strengths and vulnerabilities. Appropriate research should be conducted to develop appropriate recommendations tailored on the affordances, user bases and information practices that characterize each messaging app. Also, fact-checking operations should be continuous and adaptable, as the specifics of the misinformation phenomenon keep changing over time. Curbing misinformation online should be then embraced as a continuous effort, rather than a one-time action.

8 CONCLUSIONS

While misinformation might propagate easily on WhatsApp, the platform can also play an important role in correcting it. In this paper, we explored whether WhatsApp's affordances (flexibility in format and audience selection) and existing social capital (prevalence of strong ties and political agreement in political groups) can represent a unique opportunity for maximizing the re-share of debunking messages, such as those already accessed by WhatsApp users via existing ChatBots and Tip-Lines. We found preliminary indication that audio format might be an effective vehicle for debunks – as debunking messages received in the format of audio files generated more interest than text- or image-based messages, and they were more effective in correcting beliefs about the misinformation stimulus – although more research is needed to clarify this finding. In addition, we found clear evidence that participants re-share debunks at higher rates when they received them from people close to them (strong ties), from individuals who generally agree with them politically (in-group members), or when both conditions are met.

Rarely platforms systematically assess the efficacy of the fact-checking operations that they implement [68]. We suggest that WhatsApp leverages our findings to maximize the re-share of those fact-checks that are already circulating on the platform, as well as for future interventions. Whenever WhatsApp users access debunking messages via ChatBots and Tip-Lines, they should be actively encouraged to re-share such debunks with their contacts – most probably individuals close to them or with whom they share some similar traits – unlocking the potential for such debunks to go viral on WhatsApp. Because of the fictional elements we introduced in the design of the experiment itself, we recognise the possibility that our results might be the product of our experimental design. Our research design was constrained by the impossibility of testing our hypotheses on-platform. We look forward to conduct further experiments with the collaboration of WhatsApp, whenever possible. Ideally, future work should focus on replicating our experiments on-platform.

9 ACKNOWLEDGMENTS

EJ was supported by NSF GRFP. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation. This work was also supported by generous awards of Bill and Melinda Gates Foundation (grant number INV-010502) and Omidyar Network.

REFERENCES

[1] 2020. *Social Media and Democracy: The State of the Field, Prospects for Reform.* Cambridge University Press. https://doi.org/10.1017/9781108890960

[2] Amelia Acker. 2018. *Data Craft: The Manipulation of Social media metadata.* Data and Society. https://datasociety.net/wp-content/uploads/2018/11/DS_Data_Craft_Manipulation_of_Social_Media_Metadata.pdf

[3] Syeda Zainab Akbar, Divyanshu Kukreti, Somya Sagarika, and Joyojeet Pal. 2020. Temporal patterns in Covid 19 misinformation in India. http://joyojeet.people.si.umich.edu/temporal-patterns-in-covid-19-misinformation-in-india/.

[4] Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, Emilio Ferrara, Andrew J Flanagin, et al. 2020. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review* 1, 8 (2020).

[5] Michelle A Amazeen, Emily Thorson, Ashley Muddiman, and Lucas Graves. 2015. A comparison of correction formats: The effectiveness and effects of rating scale versus contextual corrections on misinformation. *American Press Institute. Downloaded April* 27 (2015), 2015.

[6] Michelle A. Amazeen, Chris J. Vargo, and Toby Hopp. 2019. Reinforcing attitudes in a gatewatching news era: Individual-level antecedents to sharing fact-checks on social media. *Communication Monographs* 86, 1 (Jan 2019),

112–132. https://doi.org/10.1080/03637751.2018.1521984
[7] Virginia Andreoli and Stephen Worchel. 1978. Effects of Media, Communicator, and Message Position onAttitude Change. *Public opinion quarterly* 42, 1 (1978), 59–70.
[8] Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. 1996. Identification of Causal Effects Using Instrumental Variables. *J. Amer. Statist. Assoc.* 91, 434 (1996), 444–455. https://doi.org/10.1080/01621459.1996.10476902
[9] apsa. 2019. Annenberg Presentations at APSA 2019 | Annenberg School for Communication. https://www.asc.upenn.edu/news-events/news/annenberg-presentations-apsa-2019
[10] Nahlah Ayed and Stephanie Jenzer. 2019. 'The battle is still on': Fake news rages in India's WhatsApp elections | CBC News. https://www.cbc.ca/news/world/india-whatsapp-fake-news-1.5139726
[11] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, Hong Kong, 65–74.
[12] Shakuntala Banaji, Ram Bhat, Anushi Agarwal, Nihal Passanha, and Mukti Sadhana Pravin. 2019. *WhatsApp Vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India.* London School of Economics. https://blogs.lse.ac.uk/medialse/2019/11/11/whatsapp-vigilantes-an-exploration-of-citizen-reception-and-circulation-of-whatsapp-misinformation-linked-to-mob-violence-in-india/
[13] Samarth Bansal and Poonam Snigdha. 2019. *Misinformation Is Endangering India's Election.* The Atlantic. https://www.theatlantic.com/international/archive/2019/04/india-misinformation-election-fake-news/586123/
[14] Matthew A Baum and Tim Groeling. 2009. Shot by the messenger: Partisan cues and public opinion regarding national security and war. *Political Behavior* 31, 2 (2009), 157–186.
[15] Jonah Berger. 2011. Arousal increases social transmission of information. *Psychological science* 22, 7 (2011), 891–893.
[16] Jonah Berger and Raghuram Iyengar. 2013. Communication channels and word of mouth: How the medium shapes the message. *Journal of consumer research* 40, 3 (2013), 567–579.
[17] Adam J Berinsky. 2017. Rumors and health care reform: Experiments in political misinformation. *British journal of political science* 47, 2 (2017), 241–262.
[18] Steve Booth-Butterfield and Christine Gutowski. 1993. Message modality and source credibility can interact to affect argument processing. *Communication Quarterly* 41, 1 (1993), 77–89.
[19] Danah Boyd and Alice E Marwick. 2011. Social privacy in networked publics: Teens' attitudes, practices, and strategies. In *A decade in internet time: Symposium on the dynamics of the internet and society.*
[20] Wendy J Bryce and Richard F Yalch. 1993. Hearing versus seeing: A comparison of consumer learning of spoken and pictorial information in television advertising. *Journal of Current Issues & Research in Advertising* 15, 1 (1993), 1–20.
[21] Karen E Campbell, Peter V Marsden, and Jeanne S Hurlbert. 1986. Social resources and socioeconomic status. *Social networks* 8, 1 (1986), 97–117.
[22] Karen Church and Rodrigo De Oliveira. 2013. What's up with WhatsApp? Comparing mobile instant messaging behaviors with traditional SMS. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services.* 352–361.
[23] Ann N Crigler, Marion Just, and W Russell Neuman. 1994. Interpreting visual versus audio messages in television news. *Journal of communication* 44, 4 (1994), 132–149.
[24] kyle Daly. 2021. The online far right is moving underground. https://www.axios.com/the-online-far-right-is-moving-underground-e429d45d-1b30-46e0-82a3-6e240bf44fef.html
[25] Philipe de Freitas Melo, Carolina Coimbra Vieira, Kiran Garimella, Pedro O. S. Vaz de Melo, and Fabrício Benevenuto. 2020. Can WhatsApp Counter Misinformation by Limiting Message Forwarding?. In *Complex Networks and Their Applications VIII (Studies in Computational Intelligence)*, Hocine Cherifi, Sabrina Gaito, José Fernendo Mendes, Esteban Moro, and Luis Mateus Rocha (Eds.). Springer International Publishing, Cham, 372–384. https://doi.org/10.1007/978-3-030-36687-2_31
[26] Nicholas Dias, Gordon Pennycook, and David G. Rand. 2020. *Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media.* Harvard Kennedy School Misinformation Review. https://doi.org/10.37016/mr-2020-001
[27] Rafael Evangelista and Fernanda Bruno. 2019. WhatsApp and political instability in Brazil: targeted messages and political radicalisation. *Internet Policy Review* 8, 4 (2019), 1–23. https://doi.org/10.14763/2019.4.1434 Publisher: Berlin: Alexander von Humboldt Institute for Internet and Society.
[28] Gowhar Farooq. 2018. Politics of Fake News: How WhatsApp Became a Potent Propaganda Tool in India. *Media Watch* 9, 1 (March 2018), 106–117. https://doi.org/10.15655/mw/2018/v9i1/49279
[29] Brian Feldman. 2020. Study Shows Possible Downside of Fact-Checking On Facebook. https://nymag.com/intelligencer/2020/03/study-shows-possible-downside-of-fact-checking-on-facebook.html
[30] Daniela Flamini. 2019. WhatsApp efforts to curb misinformation aren't entirely effective, research shows. https://www.poynter.org/fact-checking/2019/whatsapp-efforts-to-curb-misinformation-arent-entirely-

effective-research-shows/
[31] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.
[32] Kiran Garimella and Dean Eckles. 2020. *Images and Misinformation in Political Groups: Evidence from WhatsApp in India.* Harvard Kennedy School Misinformation Review. https://misinforeview.hks.harvard.edu/article/images-and-misinformation-in-political-groups-evidence-from-whatsapp-in-india/
[33] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press, Cambridge.
[34] Alan S Gerber and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation.* WW Norton, New York City.
[35] Hyungjin Gill and Hernando Rojas. 2020. Chatting in a mobile chamber: effects of instant messenger use on tolerance toward political misinformation among South Koreans. *Asian Journal of Communication* 0, 0 (Sept. 2020), 1–24. https://doi.org/10.1080/01292986.2020.1825757 Publisher: Routledge _eprint: https://doi.org/10.1080/01292986.2020.1825757.
[36] Vindu Goel, Suhasini Raj, and Priyadarshini Ravichandran. 2018. *How WhatsApp Leads Mobs to Murder in India (Published 2018).* The New York Times. https://www.nytimes.com/interactive/2018/07/18/technology/whatsapp-india-killings.html,https://www.nytimes.com/interactive/2018/07/18/technology/whatsapp-india-killings.html
[37] Mel Grau. 2020. *New WhatsApp chatbot unleashes power of worldwide fact-checking organizations to fight COVID-19 misinformation on the platform.* Poynter. https://www.poynter.org/fact-checking/2020/poynters-international-fact-checking-network-launches-whatsapp-chatbot-to-fight-covid-19-misinformation-leveraging-database-of-more-than-4000-hoaxes/
[38] Lucas Graves, Brendan Nyhan, and Jason Reifler. 2016. Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. *Journal of Communication* 66, 1 (2016), 102–138.
[39] Jimmeka J Guillory and Lisa Geraci. 2013. Correcting erroneous inferences in memory: The role of source credibility. *Journal of Applied Research in Memory and Cognition* 2, 4 (2013), 201–209.
[40] Aulakh Gulveen and Kalyan Parbat. 2012. Goodbye SMS? Instant messaging apps like WhatsApp & BlackBerry Messenger gaining popularity in India - The Economic Times. https://economictimes.indiatimes.com/industry/telecom/goodbye-sms-instant-messaging-apps-like-whatsapp-blackberry-messenger-gaining-popularity-in-india/articleshow/17512831.cms?from=mdr
[41] Chip Heath, Chris Bell, and Emily Sternberg. 2001. Emotional selection in memes: the case of urban legends. *Journal of personality and social psychology* 81, 6 (2001), 1028.
[42] Elizabeth E. Housholder 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 (Oct 2014), 368–382. https://doi.org/10.1080/19331681.2014.951753
[43] IANS. 2014. Why WhatsApp is so big in India- Business News. https://www.businesstoday.in/technology/news/whatsapp-facebook-trai-mobile-messaging-hotmail-microsoft/story/203524.html
[44] Jennifer Ihm and Eun-mee Kim. 2018. The hidden side of news diffusion: Understanding online news sharing as an interpersonal behavior. *New Media & Society* 20, 11 (Nov. 2018), 4346–4365. https://doi.org/10.1177/1461444818772847 Publisher: SAGE Publications.
[45] Rishi Iyengar. 2019. *WhatsApp now has a tip line for Indian election misinformation.* CNN Digital. https://www.cnn.com/2019/04/02/tech/whatsapp-india-tip-line-election/index.html
[46] Youjung Jun, Rachel Meng, and Gita Venkataramani Johar. 2017. Perceived social presence reduces fact-checking. *Proceedings of the National Academy of Sciences* 114, 23 (2017), 5976–5981.
[47] Hyunjin Kang, Keunmin Bae, Shaoke Zhang, and S Shyam Sundar. 2011. Source cues in online news: Is the proximate source more powerful than distal sources? *Journalism & Mass Communication Quarterly* 88, 4 (2011), 719–736.
[48] Evangelos Karapanos, Pedro Teixeira, and Ruben Gouveia. 2016. Need fulfillment and experiences on social media: A case on Facebook and WhatsApp. *Computers in Human Behavior* 55 (Feb. 2016), 888–897. https://doi.org/10.1016/j.chb.2015.10.015
[49] Sandhya Keelery. 2019. India - social networks used to access news 2019. https://www.statista.com/statistics/1026234/india-social-networks-used-to-access-news/
[50] Nabeela Khan. 2020. Trends in Covid-19 misinformation in India. https://www.ha-asia.com/trends-in-covid-19-misinformation-in-india/.
[51] Uttam Kumar. 2018. For PM Modi's 2019 campaign, BJP readies its WhatsApp plan. https://www.hindustantimes.com/india-news/bjp-plans-a-whatsapp-campaign-for-2019-lok-sabha-election/story-lHQBYbxwXHaChc7Akk6hcI.html
[52] Zhicong Lu, Yue Jiang, Cheng Lu, Mor Naaman, and Daniel Wigdor. 2020. The Government's Dividend: Complex Perceptions of Social Media Misinformation in China. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.

1145/3313831.3376612
[53] Arthur Lupia and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need to Know?* Cambridge University Press, Cambridge, U.K.
[54] Diane M Mackie, Leila T Worth, and Arlene G Asuncion. 1990. Processing of persuasive in-group messages. *Journal of personality and social psychology* 58, 5 (1990), 812.
[55] Harrison Mantas. 2021. Growing usage of encrypted messaging apps could make it harder to combat misinformation. https://www.poynter.org/fact-checking/2021/growing-usage-of-encrypted-messaging-apps-could-make-it-harder-to-combat-misinformation/
[56] Drew B Margolin, Aniko Hannak, and Ingmar Weber. 2018. Political fact-checking on Twitter: When do corrections have an effect? *Political Communication* 35, 2 (2018), 196–219.
[57] Alexandre Maros, Jussara Almeida, Fabrício Benevenuto, and Marisa Vasconcelos. 2020. Analyzing the Use of Audio Messages in WhatsApp Groups. In *Proceedings of The Web Conference 2020*. 3005–3011.
[58] Peter V. Marsden and Karen E. Campbell. 1984. Measuring Tie Strength. *Social Forces* 63, 2 (1984), 482–501. http://www.jstor.org/stable/2579058
[59] James C. McCroskey and Jason J. Teven. 1999. Goodwill: A reexamination of the construct and its measurement. *Communication Monographs* 66, 1 (Mar 1999), 90–103. https://doi.org/10.1080/03637759909376464
[60] Elliott McGinnies and Charles D Ward. 1980. Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin* 6, 3 (1980), 467–472.
[61] Mary Meeker. 2017. Mary Meeker Internet trends 2017 report: The top highlights for India. https://indianexpress.com/article/technology/tech-news-technology/mary-meeker-internet-trends-2017-report-india-top-highlights-points-to-note-mobile-growth-4683709/
[62] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.
[63] Microsoft. 2019. Microsoft releases digital civility index on Safer Internet Day. https://news.microsoft.com/en-in/microsoft-digital-civility-index-safer-internet-day-2019/ Section: Press Releases.
[64] Katherine L Milkman and Jonah Berger. 2014. The science of sharing and the sharing of science. *Proceedings of the National Academy of Sciences* 111, Supplement 4 (2014), 13642–13649.
[65] S U Noble. 2018. Algorithms of Oppression. https://nyupress.org/9781479837243/algorithms-of-oppression
[66] Kenton O'Hara, Michael Massimi, Richard Harper, Simon Rubens, and Jessica Morris. 2014. *Everyday Dwelling with WhatsApp*. Microsoft Research. https://www.microsoft.com/en-us/research/publication/everyday-dwelling-with-whatsapp/
[67] Priya Pathak. 2019. WhatsApp is now 10 years old and here is a look at how it grew, changed the world. https://www.indiatoday.in/technology/features/story/whatsapp-is-now-10-years-old-and-here-is-a-look-at-how-it-grew-changed-the-world-1465208-2019-02-26
[68] Gordon Pennycook and David Rand. 2020. *Opinion | The Right Way to Fight Fake News*. The New York Times. https://www.nytimes.com/2020/03/24/opinion/fake-news-social-media.html
[69] Billy Perrigo. 2019. How Whatsapp Is Fueling Fake News Ahead of India's Elections. https://time.com/5512032/whatsapp-india-election-2019/
[70] Nathaniel Persily and Joshua A Tucker. 2020. *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press.
[71] Richard E Petty and John T Cacioppo. 1979. Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of personality and social psychology* 37, 10 (1979), 1915.
[72] Michael Pfau, R Lance Holbert, Stephen J Zubric, Nilofer H Pasha, and Wei-Kuo Lin. 2000. Role and influence of communication modality in the process of resistance to persuasion. *Media Psychology* 2, 1 (2000), 1–33.
[73] Chanthika Pornpitakpan. 2004. The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence. *Journal of Applied Social Psychology* 34, 2 (2004), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x
[74] Gustavo Resende, Philipe Melo, Julio C. S. Reis, Marisa Vasconcelos, Jussara M. Almeida, and Fabrício Benevenuto. 2019. Analyzing Textual (Mis)Information Shared in WhatsApp Groups. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 225–234. https://doi.org/10.1145/3292522.3326029
[75] Richard Rogers. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication* 35, 3 (2020), 213–229.
[76] Patrícia Rossini, Jennifer Stromer-Galley, Erica Anita Baptista, and Vanessa Veiga de Oliveira. 2020. Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections. *New Media & Society* 00, 0 (June 2020), 1461444820928059. https://doi.org/10.1177/1461444820928059 Publisher: SAGE Publications.

[77] Patrícia Rossini, Jennifer Stromer-Galley, Erica Anita Baptista, and Vanessa Veiga de Oliveira. 2020. Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections. *New Media & Society* (2020), 1461444820928059.
[78] Andres Schipani, Madhumita Murgia, and Stephanie Findlay. 2019. India: the WhatsApp election. https://www.ft.com/content/9fe88fba-6c0d-11e9-a9a5-351eeaef6d84
[79] Niharika Sharma. 2020. Pesky forwards aside, WhatsApp is now a lifeline for many small businesses in India. https://qz.com/india/1882489/whatsapp-helping-indias-small-businesses-survive-covid-19-stress/
[80] Jieun Shin and Kjerstin Thorson. 2017. Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media. *Journal of Communication* 67, 2 (Apr 2017), 233–255. https://doi.org/10.1111/jcom.12284
[81] Tomer Simon, Avishay Goldberg, Dmitry Leykin, and Bruria Adini. 2016. Kidnapping WhatsApp – Rumors during the search and rescue operation of three kidnapped youth. *Computers in Human Behavior* 64 (Nov. 2016), 183–190. https://doi.org/10.1016/j.chb.2016.06.058
[82] Singh. 2019. WhatsApp reaches 400 million users in India, its biggest market. https://social.techcrunch.com/2019/07/26/whatsapp-india-users-400-million/
[83] Shivam Shankar Singh. 2019. A former BJP data analyst reveals how the party's WhatsApp groups work. https://qz.com/india/1553765/bjps-whatsapp-ops-is-what-cambridge-analytica-can-only-dream-of/
[84] Marko M Skoric, Clarice Sim, Han Teck Juan, and Pam Fang. 2009. Podcasting and politics in Singapore: an experimental study of medium effects. *Journal of Contemporary Eastern Asia* 8, 2 (2009), 27–43.
[85] With Lucile Stengel and Sapna Solanki. 2018. *Fake news and the ordinary citizen in India.* BBC.
[86] Briony Swire, Adam J Berinsky, Stephan Lewandowsky, and Ullrich KH Ecker. 2017. Processing political misinformation: comprehending the Trump phenomenon. *Royal Society open science* 4, 3 (2017), 160802.
[87] Hannah Tankovska. 2021. Most Popular Global Mobile Messenger Apps. https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/
[88] Vernise Tantuco. 2021. On Facebook's messaging apps, false information spreads undetected, unchecked. https://www.disinfo.eu/publications/on-facebooks-private-messaging-apps-harmful-misinformation-spreads-largely-undetected-and-unchecked/
[89] TechRepublic. 2013. Apple vs. Android: What does India think? https://www.techrepublic.com/blog/asian-technology/apple-vs-android-what-does-india-think/
[90] TheGuardian. 2018. 'WhatsApp murders': India struggles to combat crimes linked to messaging service. http://www.theguardian.com/world/2018/jul/03/whatsapp-murders-india-struggles-to-combat-crimes-linked-to-messaging-service Section: World news.
[91] Sebastián Valenzuela, Ingrid Bachmann, and Matías Bargsted. 2019. The Personal Is the Political? What Do WhatsApp Users Share and How It Matters for News Knowledge, Polarization and Participation in Chile. *Digital Journalism* 0, 0 (Nov. 2019), 1–21. https://doi.org/10.1080/21670811.2019.1693904 Publisher: Routledge _eprint: https://doi.org/10.1080/21670811.2019.1693904.
[92] Augusto Valeriani and Cristian Vaccari. 2018. Political talk on mobile instant messaging services: a comparative analysis of Germany, Italy, and the UK. *Information, Communication & Society* 21, 11 (Nov. 2018), 1715–1731. https://doi.org/10.1080/1369118X.2017.1350730 Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2017.1350730.
[93] Guda Van Noort, Marjolijn L Antheunis, and Eva A Van Reijmersdal. 2012. Social connections and the persuasiveness of viral campaigns in social network sites: Persuasive intent as the underlying mechanism. *Journal of Marketing Communications* 18, 1 (2012), 39–53.
[94] Emily K Vraga 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 (2018), 1337–1353.
[95] Hootsuite We Are Social. 2021. Data Report Global Overview. https://datareportal.com/reports/?tag=Global+Overview
[96] Joshua Bleiberg and Darrell M. West. 2015. Jumping from fixed Internet to mobile: India is going wireless. https://www.brookings.edu/blog/techtank/2015/03/18/jumping-from-fixed-internet-to-mobile-india-is-going-wireless/
[97] WhatsApp. 2020. Keeping WhatsApp Personal and Private. https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private/?lang=en
[98] WhatsApp. 2020. Search the Web. https://blog.whatsapp.com/search-the-web/?lang=en
[99] Andrea C Wojnicki and David Godes. 2008. Word-of-mouth as self-enhancement. *HBS marketing research paper* 01 (2008).
[100] Dannagal G Young, Kathleen Hall Jamieson, Shannon Poulsen, and Abigail Goldring. 2018. Fact-checking effectiveness as a function of format and tone: Evaluating FactCheck. org and FlackCheck. org. *Journalism & Mass Communication Quarterly* 95, 1 (2018), 49–75.

A APPENDIX

In this section, we present the same results as in section 6 but for a similar experiment conducted in Pakistan. The experiment flow in Pakistan was similar to the one explain in section 4.2 above, with the exception that the misinformation piece and its debunk message were tailored to Pakistan's political landscape and the survey was conducted in Urdu.

Of the 1457 participants recruited in Pakistan, more than half failed to pass any of the three attention check questions we had inserted at different points of the survey, and thus our final data size was restricted to $N = 646$. Figure 8 is very similar to figure 1 in the main text but it shows basic descriptive stats on demographic variables of our participants in Pakistan. In contrast to India, we were not able to survey for marital status.

A.1 H1 & H2 & H3: Audio corrections generated more interest and higher belief change than text or image-based corrections.

Figure 9 is constructed similar to figure 5 in the main text but based on our subjects in Pakistan. Figure 9a compares the levels of interests generated by each correction format. The chi-squared test of independence between audio and other formats is marginally significant ($p = 0.1$). Figure 9b compares the effectiveness of each format on correcting beliefs. The chi-squared test between three formats and belief change is marginally significant ($p = 0.055$) but the same test has enough power when we focus our attention on audio and text formats ($p = 0.016$). Table 4 shows the same results we iterated above, using ordinal logistic regression in a manner similar to table 1 in the main text. Overall, the results on debunk format in Pakistan are in agreement with those in India, albeit the dependence between format and belief change or interest is more visible in our Indian data.

A.2 H4: Corrections received by a close friend or family member are re-shared more than corrections received by a casual acquaintance

Figure 10 resembles figure 6 in the main text, but it compares re-sharing against tie strength in Pakistan. Sub-figure 10a shows the re-sharing decision over different levels of randomized tie strength treatment post-stratified on news discussion variable. Cochran–Mantel–Haenszel test with news discussion as the control category suggests randomized tie strength treatment and re-sharing are dependent with strong significance ($p = 0.008$). Sub-figure 10b compares re-sharing across different levels of self-reported tie strength after collapsing its 5 categories into 3 (Chi-squared test $p = 0.005$). In sub-figure 10c, we compare the mean re-sharing decision (after converting it to an integer ranging from 0 to 4) across the binary self-reported tie strength (t-test $p < 10^{-4}$).

Table 5 shows the same results we iterated above in terms of regression coefficients. We use the same intention to treat and encouragement design models as explained for table 2. Columns 1 and 2 use ordinal logistic regression while columns 3 and 4 use instrumental variable linear regression. Overall, the results in figure 10 and table 5 confirm our hypothesis that debunk messages received from strong ties are more likely to be re-shared. They are also in agreement with those in India shown in the main text, with two differences: (1) The baseline re-sharing decision is much lower in Pakistan than in India. (2) The results based on our random instruments are stronger in Pakistan than India. In contrast, the dependence between self-reported measures of tie-strength and re-sharing decisions is much smaller in Pakistan than in India.

A.3 H5: Corrections received by a like-minded individual are re-shared more than corrections received by an unsympathetic individual

Figure 11 resembles figure 7 in the main text, but it compares re-sharing against tie agreement in Pakistan. Sub-figure 11a shows the re-sharing decision over different levels of randomized

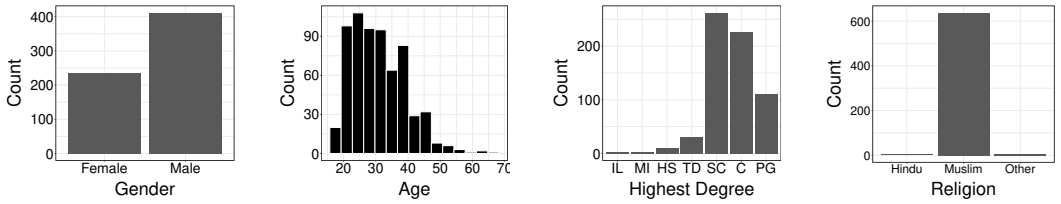
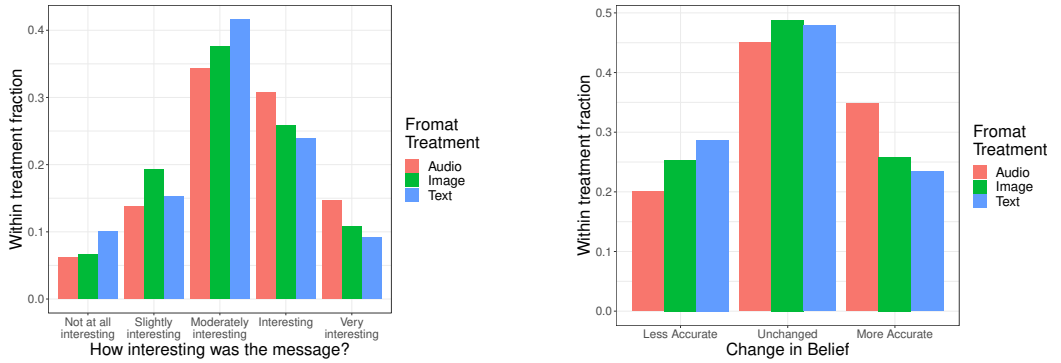


Fig. 8. The distribution of experiment participants in Pakistan across various demographic variables. In the Highest Degree subplot, IL indicates illiterate, MI indicates middle school, HS indicates High school, SC indicates Some College, TD indicates Technical Degree, C indicates College and PG indicates Post-graduate.

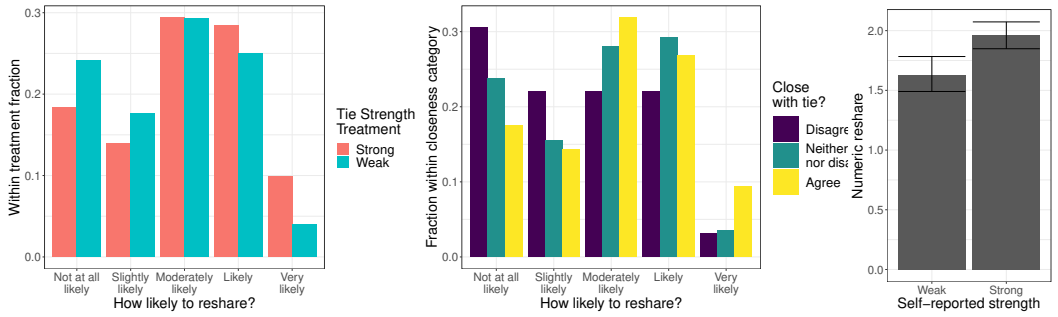


(a) (b)

Fig. 9. (a) compares how interesting the participants found the debunk message across different formats. (b) compares the level of belief change on the misinformation after observing the debunk message across different formats. For simplicity, the levels of belief change are collapsed into 3 categories.

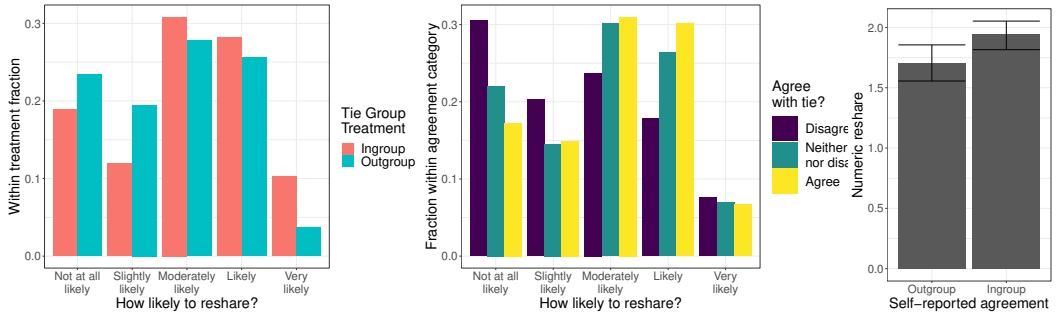
tie agreement treatment post-stratified on news discussion variable. Cochran–Mantel–Haenszel test with news discussion as the control category suggests randomized tie group treatment and re-sharing are dependent with strong significance ($p < 10^{-3}$). Sub-figure 11b compares re-sharing across different levels of self-reported tie agreement after collapsing its 5 categories into 3 similar to the main text (Chi-squared test $p = 0.04$). In sub-figure 11c, we compare the mean re-sharing decision (after converting it to an integer ranging from 0 to 4) across the binary self-reported tie agreement (t-test $p = 0.01$).

Table 6 shows the same results we iterated above in terms of regression coefficients. We use the same intention to treat and encouragement design models as explained for table 3 in the main text and table 5 for tie strength above. Overall, the results in figure 11 and table 6 confirm our hypothesis that debunk messages received from in-group ties are more likely to be re-shared. They are also in agreement with the corresponding figures from India shown in the main text. Similar to the case of tie strength, we observe that the dependence between our random tie group treatment and re-sharing decision are stronger in Pakistan than India. In contrast, the results based on self-reported measures of tie-group are much weaker in Pakistan than in India. In particular, we don't observe any significant effects in the encouragement model due to low compliance and weaker association between the self-reported measure of tie agreement and re-sharing in our Pakistan data.



(a) (b) (c)

Fig. 10. (a) compares decision to reshare the debunk message over tie strength treatments, with post-stratification using frequency of news discussion on WhatsApp. (b) compares decision to reshare the debunk messages over different levels of self-reported tie strength. The legend key is the self-reported tie strength and it corresponds to how much the participant agrees with the statement: "I am close with the tie". "Strongly (dis)agree" and "(Dis)agree" categories are collapsed into a single level. (c) compares the mean decision to reshare in numeric form (ranging from 1 to 5) between different levels of self-reported tie strength collapsed into 2 levels. Bars correspond to 95% confidence interval.



(a) (b) (c)

Fig. 11. (a) compares decision to reshare the debunk message over tie group treatments, with post-stratification using frequency of news discussion on WhatsApp. (b) compares decision to reshare the debunk messages over different levels of self-reported tie group. The legend key is the self-reported tie group and it corresponds to the participants answer to the question: "How much does the tie disagree or agree with your political beliefs?". "Strongly (dis)agree" and "(Dis)agree" categories are collapsed into a single level. (c) compares the mean decision to reshare in numeric form (ranging from 1 to 5) between different levels of self-reported tie group collapsed into 2 levels. Bars correspond to 95% confidence interval.

Received January 2021; revised July 2021; accepted November 2021

Table 4. Ordinal Logistic Regression using Randomized format treatment in Pakistan. Columns (1) and (2) use the interest in the debunk message with 5-levels as the ordinal dependent variable. Columns (3) and (4) use the change in belief about the misinfo with 11-levels as the ordinal dependent variable. Columns (2) and (4) include the categorical frequency of news discussion on WhatsApp as a control variable. The intercept corresponds to "Never Discuss News"

	Dependent variable:			
	Interest in Debunk Message		Misinfo Belief Change	
	(1)	(2)	(3)	(4)
Image Format	-0.346^{**}	-0.413^{**}	-0.370^{**}	-0.360^{**}
	p = 0.047	p = 0.019	p = 0.037	p = 0.044
Text Format	-0.461^{***}	-0.438^{**}	-0.524^{***}	-0.550^{***}
	p = 0.009	p = 0.013	p = 0.004	p = 0.003
Rarely Discuss News		0.770^{**}		-0.567^{**}
		p = 0.011		p = 0.038
Sometimes Discuss News		0.972^{***}		-0.713^{***}
		p = 0.0004		p = 0.004
Frequently Discuss News		1.764^{***}		-0.834^{***}
		p = 0.000		p = 0.003
Very frequently Discuss News		2.152^{***}		-0.686^{*}
		p = 0.000		p = 0.058
Observations	646	646	646	646
Akaike Inf. Crit.	1,891.630	1,842.857	2,023.132	2,020.358
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5. The effect of tie strength on intention to re-share the debunk message in Pakistan. Columns (1) and (2) correspond to the intention to treat causal effect, which uses the randomized tie strength assignment in an ordinal logistic regression model. Columns (3) and (4) correspond to the causal effect of tie strength using the encouragement design, where the randomized tie strength assignment is used as an instrument for the self-reported tie strength in a linear model. Note that coefficients in columns (1) and (2) are not comparable with coefficient in columns (3) and (4) since one uses logistic regression and another uses linear regression.

	Dependent variable: Re-sharing the debunk			
	Intention to Treat Model		Encouragement Design Model	
	(1)	(2)	(3)	(4)
Weak Tie	-0.469^{***}	-0.499^{***}	-2.631^{**}	-2.731^{***}
	p = 0.001	p = 0.001	p = 0.011	p = 0.010
Slightly Accurate Debunk	0.699^{***}	0.574^{**}	0.748^{***}	0.661^{**}
	p = 0.007	p = 0.029	p = 0.004	p = 0.013
Moderately Accurate Debunk	1.317^{***}	0.970^{***}	1.290^{***}	1.113^{***}
	p = 0.00000	p = 0.0001	p = 0.00001	p = 0.0003
Accurate Debunk	1.662^{***}	1.402^{***}	1.143^{***}	0.971^{***}
	p = 0.000	p = 0.00000	p = 0.00001	p = 0.0002
Very Accurate Debunk	1.398^{***}	1.187^{***}	0.796^{***}	0.658^{**}
	p = 0.0001	p = 0.001	p = 0.010	p = 0.036
Rarely Discuss News		1.126^{***}		0.637^{**}
		p = 0.0002		p = 0.011
Sometimes Discuss News		1.838^{***}		0.932^{***}
		p = 0.000		p = 0.00005
Frequently Discuss News		2.125^{***}		0.800^{***}
		p = 0.000		p = 0.006
Very frequently Discuss News		2.025^{***}		0.450
		p = 0.00000		p = 0.262
Constant			1.959^{***}	1.420^{***}
			p = 0.00000	p = 0.001
Observations	646	646	646	646
Akaike Inf. Crit.	1,920.006	1,861.022		
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6. The effect of tie group on intention to re-share the debunk message in Pakistan. Columns (1) and (2) correspond to the intention to treat causal effect, which uses the randomized tie group assignment in an ordinal logistic regression model. Columns (3) and (4) correspond to the causal effect of tie group using the encouragement design, where the randomized tie group assignment is used as an instrument for the self-reported tie group in a linear model. Note that coefficients in columns (1) and (2) are not comparable with coefficient in columns (3) and (4) since one uses logistic regression and another uses linear regression.

	Dependent variable: Re-sharing the debunk			
	Intention to Treat Model		Encouragement Design Model	
	(1)	(2)	(3)	(4)
Outgroup Tie	-0.463^{***} p = 0.002	-0.415^{***} p = 0.005	−26.067 p = 0.740	−59.509 p = 0.895
Slightly Accurate Debunk	0.679^{***} p = 0.009	0.556^{**} p = 0.034	−0.706 p = 0.851	−0.689 p = 0.935
Moderately Accurate Debunk	1.264^{***} p = 0.000	0.923^{***} p = 0.0002	0.907 p = 0.576	4.573 p = 0.881
Accurate Debunk	1.549^{***} p = 0.000	1.287^{***} p = 0.00001	−1.340 p = 0.856	−1.054 p = 0.942
Very Accurate Debunk	1.366^{***} p = 0.0001	1.152^{***} p = 0.002	−1.994 p = 0.824	−2.696 p = 0.917
Rarely Discuss News		1.036^{***} p = 0.0005		−8.074 p = 0.902
Sometimes Discuss News		1.757^{***} p = 0.000		−13.333 p = 0.902
Frequently Discuss News		2.064^{***} p = 0.000		−15.928 p = 0.902
Very frequently Discuss News		1.955^{***} p = 0.000		−16.545 p = 0.901
Constant			14.215 p = 0.718	40.088 p = 0.893
Observations	646	646	646	646
Akaike Inf. Crit.	1,920.390	1,864.754		
Note:				*p<0.1; **p<0.05; ***p<0.01