Gamified Inoculation Against Misinformation in India: A Randomized Control Trial

Trisha Harjani[1], Melisa-Sinem Basol[1], Jon Roozenbeek[1], Sander van der Linden[1]

Although the spread of misinformation is a pervasive and disruptive global problem, extant research is skewed towards "WEIRD" countries leaving questions about how to tackle misinformation in the developing world with different media and consumption patterns unanswered. We report the results of a game-based intervention against misinformation in India. The game is based on the mechanism of psychological inoculation; borrowed from the medical context, inoculation interventions aim to pre-emptively neutralize falsehoods and help audiences spot and resist misinformation strategies. Though the efficacy of these games has been repeatedly demonstrated in samples from Western countries, the present study conducted in north India (n = 757) did not replicate earlier findings. We found no significant impact of the intervention on the perceived reliability of messages containing misinformation, confidence judgments, and willingness to share information with others. Our experience presents a teachable moment for the unique challenges associated with complex cultural adaptations and field work in rural areas. These results have significant ramifications for designing misinformation interventions in developing countries where misinformation is largely spread via encrypted messaging applications such as WhatsApp. Our findings contribute to the small but growing body of work looking at how to adapt misinformation interventions to cross-cultural settings.

Keywords *misinformation, India, inoculation theory, pre-bunking, WhatsApp*

[1] University of Cambridge

Received
June 7, 2022
Accepted
October 10, 2022
Published
February 27th, 2023

Correspondence
University of Cambridge
th649@cam.ac.uk

The spread of misinformation online is widely documented as a threat to democracies worldwide (Lewandowsky et al., **2017**; van der Linden, Maibach, et al., **2017**). In India, the world's largest democracy, the sharing of misinformation online has been linked to mob violence, and even killings (Arun, **2019**; Sundar et al., **2021**; Vasudeva & Barkdull, **2020**). While social media platforms such as Facebook or Twitter can flag misinformed content or remove it from their platforms, mobile instant messenger services such as WhatsApp and Telegram are limited by their end-to-end encrypted nature (Banaji et al., **2019**). Private conversations or groups form a closed network where misinformation can freely circulate without monitoring and studies have shown that this takes place in India (Badrinathan, **2021**), as well as Burundi (Mumo, **2021**), Nigeria, Brazil, and Pakistan (Pasquetto et al., **2020**). Furthermore, a significant proportion of the misinformation shared in India continues to be shared and circulated on WhatsApp even after being falsified by professional, third-party fact checkers (Reis et al., **2020**). This trend has created a breeding ground for unverified, misleading, or false information, some of which originates from political parties (Chibber & Verma, **2018**). Despite WhatsApp's countermeasures, which include implementing digital literacy programs, placing restrictions on forwarding, and broadcasting awareness-raising adverts, misinformation on the platform is persistent and has been exacerbated by COVID-19 (Al-Zaman, **2021**; Ferrara, **2020**). Given the limitations of implementing algorithmic solutions on private messaging platforms (Reis et al., **2020**), user-level solutions are an increasingly important avenue of research.

The overwhelming majority of individual-

Take-home Message

This study found that gamified inoculation interventions, which have worked well in Western countries, did not confer psychological resistance against misinformation to participants in India. This null result (possibly due to lower digital literacy rates) calls for further investigation into bottom-up interventions tackling misinformation on messaging platforms in developing countries.

level misinformation interventions have been tested on populations from developed, Western countries. This is indeed a feature of behavioral science in general where non-WEIRD (western, educated, industrialized, rich and democratic) samples are underrepresented (Henrich et al., **2010**; Rad et al., **2018**). There are several factors that could impede the generalizability of findings to India specifically. Since 2017, year-on-year internet penetration in India has grown by 13% in rural areas compared to 4% in urban neighborhoods (Bhattacharjee et al., **2021**). While misinformation can be spread by both urban and rural residents, the latter are likely to access the internet via 2G networks with limited resources for fact checking and a tendency to distribute WhatsApp messages with low reflexivity, as a mode of group participation or strategy to avoid feelings of exclusion (Banaji et al., **2019**). Given the collectivist culture in India (Kapoor et al., **2003**; Verma & Triandis, **2020**), even amongst youth samples (Rao et al., **2013**), the importance of group identities is heightened. Political parties frequently capitalize on these divisions, often along religious lines (Vaishnav et al., **2019**). Furthermore, the institutionalization of misinformation dissemination by political parties in India, whereby 'IT cells' troll and spread automated content, is not uncommon (Campbell-Smith & Bradshaw, **2019**) as part of their campaigning strategy (Banaji et al., **2019**).

To counter the spread of misinformation, several strategies have been researched at the individual level, the most well-known of which include fact-checking and "debunking" or correcting false information after exposure (Ecker et al., **2022**; van der Linden, **2022**; Walter & Murphy, **2018**). Studies examining the efficacy of such corrective measures have revealed mixed results. Although some have found that fact-checking can improve accuracy assessments (Clayton et al., **2020**; Porter & Wood, **2021**; Walter & Murphy, **2018**), there are several drawbacks to correcting misinformation post-exposure. One major issue concerns the continued influence of misinformation or the tendency for people to continue making inferences based on misinformation. They do so even when they acknowledge a correction (Ecker et al., **2022**; Lewandowsky et al., **2012**), which limits the correction's potential effectiveness. This is further compounded by the finding that (a) not all audiences are receptive to fact-checks (Walter et al., **2020**), (b) repeated exposure to misinformation can increase its perceived accuracy (Pennycook et al., **2018**; Swire et al., **2017**), and (c) that corrections do not scale, meaning they rarely reach the same number people as the initial misinformation (Roozenbeek & van der Linden, **2019**; van der Linden, **2022**). Lastly, corrective strategies are also difficult to implement on private messaging platforms given the invisibility of information flow in this sphere (Reis et al., **2020**).

Accordingly, studies which have evaluated fact-checking and literacy interventions in developing countries have revealed inconclusive results. For example, Guess et al. (**2020**) tested the effect of providing U. S. and Indian participants with tips on how to spot misinformation. They found a positive impact on people's ability to detect false information in the U. S. and in a highly educated online Indian sample, but not in a face-to-face sample obtained in rural Northern India. Similarly, Badrinathan (**2021**) tested the impact of an intensive one-hour in-person media literacy training during the 2019 national election and found no significant beneficial effects.

One study tested the impact of a debunking intervention via WhatsApp broadcast messaging in Zimbabwe, another country with high WhatsApp usage, finding that participants had increased knowledge about COVID-19 (Bowles et al., **2020**). Pasquetto et al. (**2020**) further found that, while corrections on encrypted group chats reduced belief in misinformation in India and Pakistan, WhatsApp users report corrections as unusual and socially awkward. Given the known challenges surrounding debunking and fact-checking, a promising ef-

fort against misinformation has been to pre-emptively debunk (or *prebunk*) falsehoods to allow individuals to acquire skills to detect and resist misinformation in the future (Lewandowsky & van der Linden, **2021**). This approach is based on the theory of psychological inoculation (McGuire, **1961**).

Theoretical Background: Prebunking and Inoculation Theory

Inoculation theory was originally developed in the 1960s and is based on the biological process of immunization (McGuire, **1961**, **1964**): just as exposure to a weakened dose of a pathogen can confer immunity against future infection(s), pre-emptively exposing people to weakened doses of misinformation—along with strong refutations—can cultivate cognitive immunity to future manipulation attempts. Inoculation theory has two key components. Firstly, the inoculation must have a forewarning to evoke threat or the motivation for people to defend themselves from a potential attack on their attitudes (Compton, **2012**). Being aware of one's vulnerability to manipulation is important for kick-starting resistance to persuasion (Sagarin et al., **2002**). Secondly, much like the injection of a weakened dose of a virus can build immunity through the production of antibodies, exposure to a weakened version of a persuasive argument along with a counterargument can inspire lowered vulnerability to misleading persuasion attempts (McGuire, **1961**). A meta-analysis of inoculation theory has found that it is effective at building resistance against persuasion across issues (Banas & Rains, **2010**).

In more recent years, the theory has informed the design of inoculation interventions aiming to endow attitudinal resistance against online misinformation specifically (for in-depth reviews see Compton et al., **2021**; Lewandowsky & van der Linden, **2021**; Roozenbeek & van der Linden, **2018**; van der Linden, **2022**). Some recent applications of inoculation theory include even potentially polarizing topics such as climate change (van der Linden, Leiserowitz, et al., **2017**), conspiracy theories (Banas & Miller, **2013**), or vaccinations (Jolley & Douglas, **2017**). However, all these studies aimed to inoculate people against misinformation about a specific issue. As such, they do not necessarily imply that the inoculation would be effective as a "broad-spectrum vaccine" against misinformation (Roozenbeek & van der Linden, **2018**). This prompted a shift away from narrow-spectrum inoculations to those that incorporate persuasion techniques common to misinformation more generally (Cook et al., **2017**; Roozenbeek & van der Linden, **2019**). In other words, familiarity with a weakened dose of the underlying *techniques* that are used to spread misinformation could impart an increased cognitive ability to detect manipulative information that makes use of such misinformation tactics. These tactics include emotionally manipulative language, group polarization, conspiratorial reasoning, trolling, and impersonations of fake experts, politicians, and celebrities (Roozenbeek & van der Linden, **2019**).

This strategy has demonstrated fairly consistent success (Basol et al., **2020**; Cook et al., **2017**; Roozenbeek & van der Linden, **2019**) including long-term efficacy, provided inoculated individuals are given short reminders or "booster shots" of the lessons learned (Maertens et al., **2021**). Yet, no study to date has tested the effect of inoculation interventions on the Indian population and inoculation researchers have noted a lack of generalizability of inoculation scholarship to non-WEIRD populations (Bonetto et al., **2018**), demanding interventions be adapted and evaluated.

Recent Applications: Inoculation Games

Recent applications of inoculation theory also depart from the traditional method of providing participants with ready-made counterarguments (so-called "passive inoculation") and instead use an "active" form of inoculation whereby participants themselves play an active role in generating resistance to manipulation (Roozenbeek & van der Linden, **2018**). Gamified interventions have proven to be a fruitful vehicle for active inoculation. One example of such an inoculation intervention is the online game *Bad News* (www.getbadnews.com): in this game, players find themselves in an artificial social media environment designed to mimic the features of widely used online platforms (Basol et al., **2020**; Maertens et al., **2021**; Roozenbeek & van der Linden, **2019**; Roozenbeek et al., **2021**). Across six levels, players

are warned about the dangers of fake news, and they develop an understanding of several widely used misinformation techniques through exposure to weakened dose of these tactics alongside ways to spot them. Evidence for the relative benefits of "active" inoculation is emerging (Basol et al., **2021**), particularly because it may strengthen associative memory networks, contributing towards higher resistance to persuasion (Pfau et al., **2005**).

However, the *Bad News* game, as well as two others (*Harmony Square* Roozenbeek & van der Linden, **2020**) and (*Go Viral!* Basol et al., **2021**), all focus on misinformation on public social media platforms (such as Facebook and Twitter). This reduces the potential applicability of these games in countries where direct messaging apps are a more common means of communication than public social media platforms. To address this problem, we engaged in a novel real-world collaboration with WhatsApp, Inc (Meta platforms) and developed a new game that inoculates people against misinformation on direct messaging apps, called *Join this Group* (link to English version; https://whatsapp.aboutbadnews.com). The Hindi-version of the game was tested in this study (further details in the method section). Its purpose is to inoculate participants against four manipulation techniques commonly present in misinformation on direct messaging apps. Specifically, these techniques are the impersonation of a fake expert (Goga et al., **2015**; Jung, **2011**; Reznik, **2013**), use of emotional language to frame content (Gross & Ambrosio, **2004**; Konijn, **2012**; Zollo et al., **2015**), polarization of narratives to create hostility towards the opposition (Groenendyk, **2018**; S. Iyengar & Krupenkin, **2018**), and the escalation of an issue such that misinformation triggers offline acts of aggression (BBC Monitoring, **2021**; Robb, **2021**).

The Present Research

This paper seeks to address two gaps in the literature on misinformation interventions. We first aim to understand whether inoculation against misinformation can improve people's ability to spot misinformation that is commonly shared in a private messaging context (such as on WhatsApp). Second, our sample is from India, an understudied population where the spread of misinformation via private messaging platforms has been linked to violence (McLaughlin, **2018**). We ran a field experiment in India testing the efficacy of the inoculation game, *Join this Group*.

This paper therefore makes two unique advancements to the literature. This study is the first to test an inoculation intervention against misinformation shared in the context of private messaging. This domain of information exchange is markedly different to public platforms such that the burden of identifying, addressing, and correcting misinformation falls on the user(s) (Pasquetto et al., **2020**). Moreover, we test the effectiveness of these modified interventions in India (n = 757), the largest market for WhatsApp globally (Findlay, **2019**). Both studies were approved by the Cambridge Psychology Research Ethics Committee (REC-2018-19/19). [Data and scripts are deposited on the Open Science Framework: https://osf.io/abjrg].

Method

We conducted a 2 (treatment – control) x 2 (pre – post) mixed-between randomized control trial on a sample collected from 8 North Indian states (Bihar, Chhattisgarh, Haryana, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh, and National Capital Territory (Delhi)). Participants were recruited as part of media literacy workshops administered to 1283 individuals. The experiment was conducted door-to-door, in person, with the assistance of iPads and smartphones through which participants could access the online intervention. After providing informed consent, participants were asked to indicate their frequency of WhatsApp usage in the last twelve months on a 5-point scale, ranging from "*Never*" to "*More than once a day*". Participants were then shown 16 screenshots of WhatsApp conversations in a randomized order (see Figure 1) and, following Roozenbeek et al. (**2021**), were asked to make three assessments: how reliable they found the post (1), how confident they are in their reliability assessment (2) and how likely they would be to share the message (3). All three assessments were rated on a 1-7 Likert scale (1 being "*Not at all*", 4 being "*Neutral*", and 7 being "*Very much*"). Of the 16 images, four were screenshots of authentic WhatsApp conversations, of which two

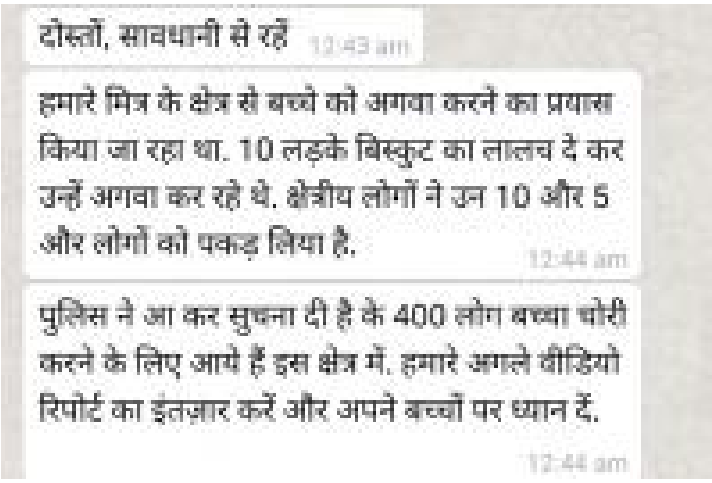


Figure 1 WhatsApp messages containing emotional misinformation messaging. This image is an example of one used in the experimental pre-test and post-test measure. The screenshot reads: "Friends, be careful", "Attempts are being made to kidnap a child from our friend's area. 10 boys were kidnapping him with the promise of biscuits. People in the area have caught those 10 and 5 more people", "The police has announced that 400 people had come to steal the child in this area. Wait for our next video that will report this and watch over your children carefully."

were fake news and two contained accurate information. The remaining 12 were screenshots containing misinformation designed to demonstrate four manipulation techniques (fake expert, emotion, polarization, and escalation). The four real (non-misinformation) items were sourced from fact-checking websites and the manipulative items were created by one of the authors and validated by two other authors, to ensure that the conversations make appropriate use of a misinformation technique. Figure 1 demonstrates an example of eliciting fear using emotional language in misinformation messaging.

Participants were then randomly assigned to play either *Join this Group* (treatment) or *Tetris* (control), consistent with previous gamified inoculation experiments (Basol et al., **2020**; Roozenbeek & van der Linden, **2020**). Gameplay for *Join this Group* was approximately 15 minutes while *Tetris* participants had to play for a minimum of nine minutes before proceeding. Participants who played *Join this Group* were required to input a password to validate their completion. Following the game, as part of the post-test measure, all participants were asked to assess the same 16 WhatsApp conversations again and answer some demographic questions, including district, state, gender, education level, age group, how frequently they check the news, how frequently they use social media platforms, their interest in politics, their political ideology, and attitudes scales assessing left to right and libertarian to authoritarian views (Park et al., **2013**). Participants were also asked to provide their first thoughts upon hearing the term "fake news."

Treatment Game: Join this Group

We created a Hindi translation of the *Join this Group* game in collaboration with a Delhi-based non-profit, the Digital Empowerment Foundation (DEF). One major challenge that arose during field implementation is that our novel inoculation approach did not fit conceptually into DEF's media literacy strategy. As a condition of administering the intervention in rural India, DEF therefore required that we adapt the intervention to be more in line with their own media literacy strategy. As a result, the key difference between the English and Hindi versions of the *Join this Group* game is that players take on more of a traditional fact-checking role by posing as an undercover detective fighting misinformation online. This is in stark contrast to active inoculation games such as *Bad News*, *GoViral!*, and *Harmony Square.* In these games, participants generally take on the role of a misinformation *spreader* because this perspective-taking exercise helps elicit "motivational threat" or the motivation to defend oneself against misinformation, a key component of inoculation theory (Basol et al., **2021**). However, DEF advised that such a perspective was not in line with their traditional media literacy training and may be confusing for their target audience in India, who generally have low digital literacy. Accordingly, we created a new version of the game where the player steps into the shoes of a fake news "detective."

In the Hindi version, players are introduced to the game with a messaging-interface screen reading "Hello detective! We need you." The game explains that a group called "Big News" is spreading propaganda on WhatsApp in the fictional nation of "Santhala." The game then explains that understanding the techniques

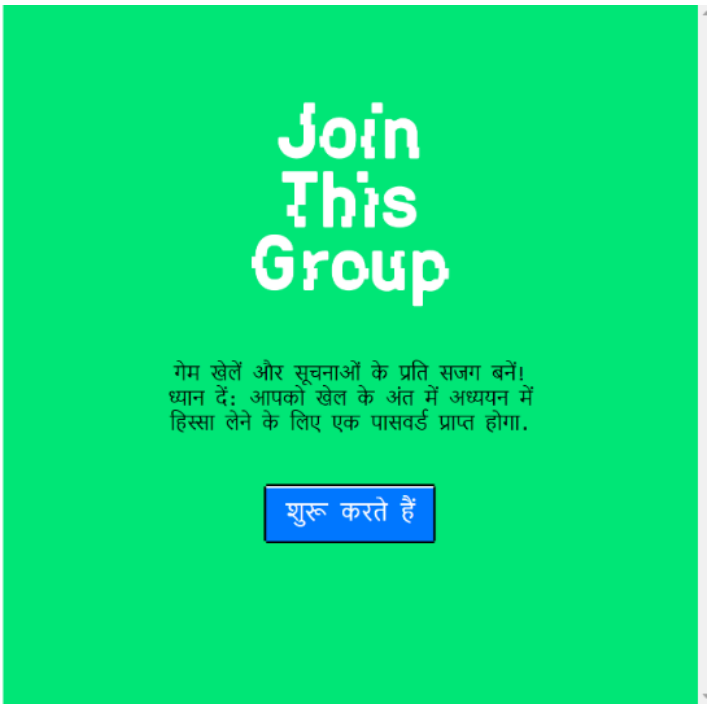


Figure 2 Landing page of the game. The text reads "Play the game and watch out for notifications! Attention: You will receive a password at the end of the game. In order to take part in the study, you'll need to input this password." Blue button reads "Let's start."

of the "Big News" group will require going undercover since messages are encrypted and untraceable. Figures 2 and 3 below display in-game screenshots. See Figures S4-S8[1] for more screenshots.

Players go through four levels, each one teaching and testing the application of techniques present in misinformation (fake experts, emotional language, polarization, escalation). See Table 1 for an overview of the four levels. In the first level, players are shown how sharing messages in a group unannounced can result in being reported, an issue that can be overcome by impersonating a fake expert to boost credibility of spurious claims. Players are then able to go undercover by spreading rumors such as "Mangoes cause cancer" using their fake pseudonym (See Figure 3). Such impersonations are pervasive throughout social media (Adewole et al., **2017**; Goga et al., **2015**; Jung, **2011**; Reznik, **2013**). The second level shows players how the use of emotionally charged language can create an atmosphere of chaos especially when combined with a visual prompt. Emotional framing and language have been shown to increase salience, social media engagement (Rathje et al., **2021**), grab attention (Konijn, **2012**), and evoke emotional reactions (Gross & Ambrosio, **2004**). The third level continues in context where players now need to apply their detective skills to prevent election manipulation. They are shown how repeated false messaging that uses partisan misinformation can vilify and antagonize the opposition (such as a political party), exaggerate the perceived distance between identities, sow doubt and increase support for a particular group (Groenendyk, **2018**; S. Iyengar & Krupenkin, **2018**; Melki & Pickering, **2014**). Finally, in the fourth level players are told that they need to report the partisan misinformation being shared. This results in the suspicion of a disloyal supporter in the political party's WhatsApp group and motivates a targeted offline attack on the mole, which intensifies into protests and riots. Throughout this level, the game explains how online encouragement can escalate into offline aggression (BBC Monitoring, **2021**; Robb, **2021**).

At the end of each level, players are given a summary of the techniques they have been inoculated against. Points and sanctions are also counted throughout; if players send a message that does not reflect use of the techniques learned, they are penalized. Conversely, exposing propaganda as an undercover detective increases points. In all scenarios, players also see WhatsApp group members' reactions to the misinformation. Overall, the game aims to demonstrate how fabricated content can evoke not only belief in misinformation but also create an atmosphere of fear, polarization, and elicit violent offline behavior.

The study was thus designed to test the efficacy of *Join this Group*, measured by three forms of assessment. We therefore hypothesized that:

H_1 Treatment group participants find manipulative WhatsApp messages significantly less reliable post-gameplay compared to the control group.

H_2 Treatment group participants are significantly more confident at assessing the reliabil-

[1] All figures and tables starting with S are to be found in the supplementary materials.

Figure 3 The first two messages after starting the game.
The top message reads "Hello Detective! We need you." The bottom message reads "Our great country Santhala needs you. A group called 'Big News' is spreading propaganda on a very large scale" (left).
In-game screenshot from the first level. The top message reads "Well done! Find the profile of a person who is a fake doctor." The bottom message reads "Dr. Saurav Agrawal" (right).

ity of manipulative WhatsApp messages compared to the control group.

H_3 Treatment group participants are significantly less likely to want to forward manipulative WhatsApp messages to others compared to the control group.

Sample

After providing informed consent, we collected n = 1283 observations, of which, n = 757 were complete responses. Participants did not always complete the full survey; we saw some drop-off after the intervention as many participants did not complete the post-test. To understand if the data was missing at random (MAR), we ran further analyses using the pre-test scores, condition allocation and WhatsApp usage data to assess missingness (see the supplementary materials for full details). We were not able to study the demographic predictors of the incomplete data because this was collected at the end of the study. The analysis finds that the data was not missing at random and that a higher baseline confidence in assessing the reliability of manipulative items decreased the odds of missingness (OR = 0.030, [95%CI; 0.002,0.431]) and being assigned to the treatment group increased the odds of missingness (OR = 2.171, [95%CI; 1.589, 2.967]). Please see Table S1 for full results.

During the data quality check, we further observed data in which participants just provided the same scale point consistently throughout the pre-test, post-test, or both (e.g., "4"). We therefore removed any responses which had repeated answer patterns[2] throughout the entire section (pre-test or post-test), resulting in a final sample size of n = 725. Of the final sample, 55% identified as female, 40% as male and 5% as other. 49% reported being 18-24 years old. 42% reported having obtained at least a bachelor's degree. The sample was also heavily left leaning, (M = 2.14, SD = 0.78). Finally, 65% of participants came from the state of Madhya Pradesh (17% from Rajasthan, 6% from Chhattisgarh, 5% from Uttar Pradesh, 4% from Jharkhand, 3% from Bihar). See Table S2 for a full breakdown of the sample.

Results

All data cleaning and analysis was conducted using RStudio, scripts are available via the Open Science Framework: https://osf.io/abjrg. For the main analyses, the following packages were used: stats (for ANCOVA), TOSTER (for tests of statistical equivalence) and BayesFactor (for Bayesian t-tests).

We conducted a one-way ANCOVA to test **H_1**,

[2]Analysis including the excluded 32 responses was also run and these did not affect the results.

Table 1 A summary of the game from the player's perspective at each of the four levels.

Level	*Manipulation Technique*	*Description*
1	Fake Expert	As undercover detectives, players join a WhatsApp group called "Breaking News" in the town of "Santhala." They share a fake message but are kicked out of the group, upon which they are encouraged to use a fake expert to gain credibility and witness how this impersonation can garner belief.
2	Emotional Language	Players are told that certain users in the group "Big News" are picking fights. As an undercover detective, they are tasked with spreading content to contribute to the chaos. The game then prompts players to share a fear or anger inducing message. This level shows players how, especially when paired with an image, emotional language can manipulate opinions and exacerbate chaos in the group.
3	Polarization	At this stage, Santhala is facing an election that the group "Breaking News" is attempting to manipulate. Players are told they must go undercover in one of the political candidate groups to spread polarizing information (e.g., damaging information about the opposition). The game shows how this cycle causes wider rifts between supporters.
4	Escalation	Continuing in context, the opposition group reports the polarizing fake news shared earlier to the media. The player is shown how members of the group try to identify the 'mole' which escalates into an offline attack on the suspected individual. Although WhatsApp now bans this political group, players are shown how they simply create another one with new phone numbers.

examining whether post-test reliability scores of manipulative items were significantly different between conditions, controlling for pre-test scores. We found no significant difference in reliability assessments between treatment and condition groups: $F(1,722) = 0.00$, $p = 0.97$. This relationship held for the subcategories of the fake items; fake expert: $F(1,722) = 0.21$, $p = 0.65$; emotion: $F(1,722) = 0.21$, $p = 0.65$; polarization: $F(1,722) = 0.35$, $p = 0.55$; and escalation: $F(1,722) = 0.03$, $p = 0.85$. To test whether the non-significant results imply null effects or equivalence to zero (Lakens et al., **2018**), we conducted an equivalence test using two one-sided tests (TOST) on the post-gameplay outcomes (TOSTs).[3] We could not confirm statistical equivalence to zero for the average reliability score $t(721.68) = -1.44$, $p = 0.07$. However, a Bayesian paired samples *t*-test for the averaged reliability score of misinformation items gives a Bayes factor of $BF_{10} = 0.25$ (error % = 0.00)**,** indicating support for the null hypothesis of $\mathbf{H_1}$ (Dienes, **2014**).

To test $\mathbf{H_2}$**,** we followed the same analysis: we conducted a one-way ANCOVA on the average post-test confidence in reliability judgment scores, controlling for the baseline. We find no significant difference between groups: $F(1,722) = 1.79$, $p = 0.18$ or for the subcategories; fake expert: $F(1, 722) = 1.56$, $p = 0.21$; emotion: $F(1,722) = 1.05$, $p = 0.31$; polarization: $F(1,722) = 1.18$, $p = 0.28$; escalation: $F(1,722) = 1.17$, $p = 0.28$. A TOST equivalence test confirmed equivalence to zero for the average post-test confidence scores (in assessing the reliability of misinformation items), $t(721.43) = -2.34$, $p = 0.01$. A Bayesian *t*-test provided strong evidence for the null hypothesis of $\mathbf{H_2}$, with a Bayes factor of $BF_{10} = 0.04$ (error % = 0.00).

To test $\mathbf{H_3}$, or whether there was a difference in post-test scores of intended willingness to share misinformation, another one-way ANCOVA was conducted on the average post-test scores, controlling for the baseline. Results were non-significant $F(1,722) = 1.46$, $p = 0.23$ including on the subcategories; fake expert: $F(1,722) = 1.94$, $p = 0.16$; emotion: $F(1,722) = 0.29$, $p = 0.59$; polarization: $F(1,722) = 2.75$, $p = 0.10$; and escalation: $F(1,722) = 2.77$, $p = 0.10$. A TOST analysis on the post-test likelihood to

[3]The smallest effect size of interest (SESOI) was set to d = ±0.25 based on the smallest observed effect size found in published experiments that use gamified inoculation interventions (Roozenbeek & van der Linden, **2019**).

share misinformation items scores could not confirm statistical equivalence to zero $t(719.73) = -0.64$, $p = 0.26$. However, a Bayesian t-test suggested strong support for the null hypothesis of $\mathbf{H_3}$ with a Bayes factor of $BF_{10} = 0.07$ (error % = 0.00). See Table S6 for Bayesian t-tests. Figure 4 shows the distribution of mean scores (reliability, confidence and sharing) for all misinformation items. Similarly, Figure 5 displays the distribution of mean reliability scores broken down by technique.

Though not hypothesized, to test whether the intervention increased skepticism towards factual messages, we also conducted a one-way ANCOVA to test for significant differences in post-gameplay scores for real news items, controlling for baseline scores. Specifically, ratings of reliability: $F(1,722) = 0.09$, $p = 0.76$; confidence in judgments: $F(1,722) = 1.10$, $p = 0.30$; and likelihood to share: $F(1,722) = 1.39$, $p = 0.24$ were not significantly different across treatment and control groups. Similarly, we tested whether the intervention improved participants assessments of the two genuine screenshots capturing fake news sharing on WhatsApp. Using one-way ANCOVAs we found no significant differences in ratings of reliability: $F(1, 712) = 0.99$, $p = 0.32$; confidence: $F(1, 711) = 1.68$, $p = 0.20$; or likelihood to share: $F(1,702) = 0.12$, $p = 0.73$.

We ran linear regressions to check for covariate effects on the differences in pre-post measures of reliability, confidence, and sharing. We only find that higher frequency of checking the news significantly predicts a larger difference between pre and post confidence scores of misinformation items ($p = 0.03$). See Tables S33-S35 for the full results.

Discussion

Through this study we demonstrate that there was no significant effect of playing *Join this Group* on the veracity evaluations of both real and misinformation items in our sample of North Indians. This is in contrast with previous results that have found promising results using gamified inoculation in Western populations (including versions translated to German, Greek, French, Polish, and Swedish (Basol et al., **2021**; Roozenbeek & van der Linden, **2020**). Direct replications of the *Bad News* game online have also shown positive effects on urban populations in India (A. Iyengar et al., **2022**) and importantly, randomized trial data[4] from a representative sample of the UK population using the English version of *Join this Group* found that the game significantly improved people's ability to detect fake news, how confident they were in their own judgments, and reduced their overall willingness to share misinformation with others (Basol et al., **2022**).

There could be a myriad of explanations for the discrepant results observed. therefore, we categorize explanations into two broad categories: (1) cross-cultural (Indian sample, translated to Hindi) and (2) perspective shift (the player assumed the role of detective).

Firstly, we discuss possible cross-cultural explanations for our observed findings. While inoculation interventions demonstrate a clear potential to be effective (Traberg et al., **2022**), it is not surprising that the process of applying an intervention to understudied, non-WEIRD cultures (Henrich et al., **2010**; Rad et al., **2018**) might require an iterative process. Indeed, previous interventions aiming to reduce belief in and sharing of misinformation in India have faced similar difficulty. WhatsApp's media literacy campaigns and adverts have been criticized for a lack of alignment with local contexts (Medeiros & Singh, **2021**). In-person or online digital literacy interventions have either demonstrated no reduced belief in misinformation (Badrinathan, **2021**) or an effect size limited to a highly educated subset (Guess et al., **2020**). Here, we tested the efficacy of an inoculation intervention, *Join this group,* that was modified for context through partnership with a local non-profit. The intervention aimed to teach participants fundamental techniques commonly used in the presentation of misinformation through an inoculation intervention. We expected that our local adaptation and use of inoculation would improve individual veracity discernment of manipulative news items. Yet, we do not find this in our study.

We hypothesize that the cultural context, local values, and social preferences may have played a role. In particular, the process of successful inoculation in the Indian population may be different. Threat has long been conceptualized a key and necessary component for inoculation to take place (McGuire,

[4]This publication of this data is forthcoming. Once published, it can be made available upon request.

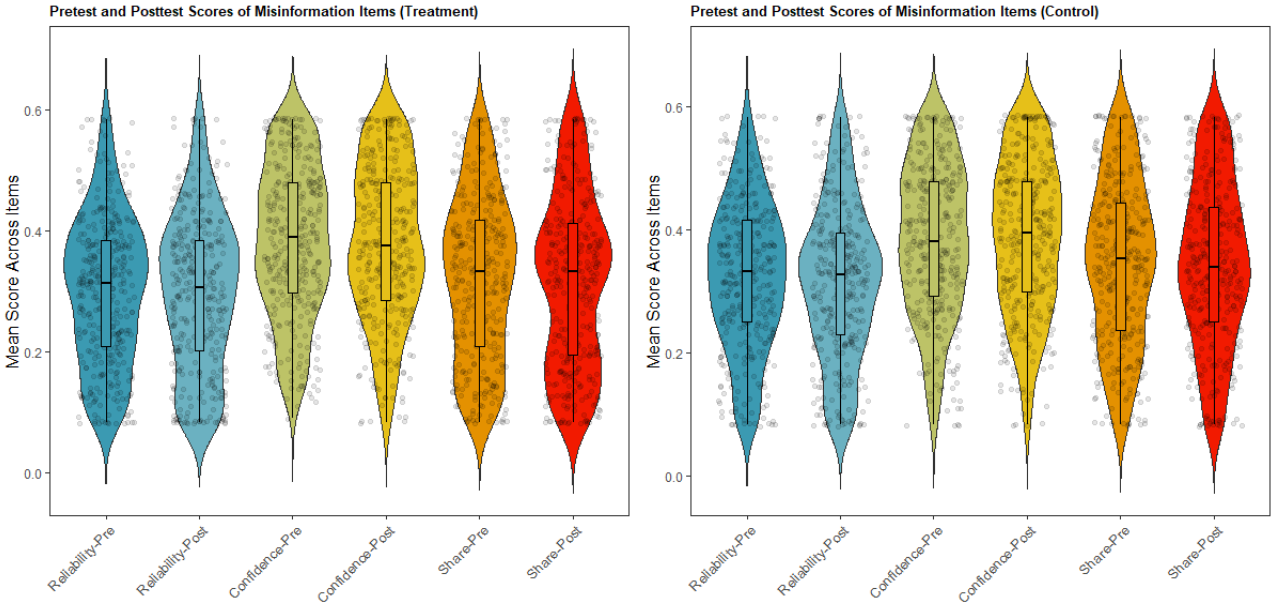


Figure 4 Distribution of pre-test and post-test mean scores in the treatment and control groups, for the reliability, confidence, and sharing scores of misinformation items across all manipulation techniques.

1964) with most recent scholars agreeing that a threshold level of threat is required for inoculation to be conferred (Compton, **2021**) as it serves the function of highlighting one's vulnerability which in turn, motivates the build-up of resistance. While there is no quantitatively defined level of minimum threat discussed in inoculation theory, studies assessing inoculation have traditionally measured threat as an apprehension (Ivanov et al., **2022**; Wood, **2007**) and more recently in a motivational form (Banas & Richards, **2017**). Unfortunately, we did not include measures of apprehensive or motivational threat in our study. Moreover, given the paucity of literature around non-WEIRD samples in psychology in general, it is difficult to make claims about the efficacy of inoculation without an explicit measurement of threat. Future research should consider incorporating this, informed by cultural variation in emotional experience and motivations (Kwan, **2016**; Lim, **2004**; Matsumoto et al., **2008**; Mesquita & Walker, **2003**).

The cross-cultural adaptation also required numerous language and context changes. (Roozenbeek & van der Linden, **2020**). For example, the chosen fictional country of "Santhala" may have carried pre-conceived notions for some given its close resemblance to the Santhal tribe (The Editors of Encyclopaedia Britannica, **2012**). All 12 manipulative WhatsApp prompts were translated from English to Hindi, which may have resulted in a loss of meaning and validity of measurement (see Figure S9 for an example). In addition, based on 2011 national census data, we estimate that our sample is 74% rural (Government of India, **2016**), a figure calculated based on the sample's distribution across states (see Table S39). Shahid et al. (**2022**) find that rural samples had a lower ability to detect misinformation compared to their urban counterparts, suggesting that interventions on rural samples may require additional challenges.

Moreover, rural areas are estimated to have a digital literacy rate of 25% compared to 61% in urban areas (Mothkoor & Mumtaz, **2021**), suggesting that our sample has low digital literacy overall. Classifying a household as digitally literate only requires one person, aged above 5 years, to be able to operate a computer and use the internet. As such, it is likely that our

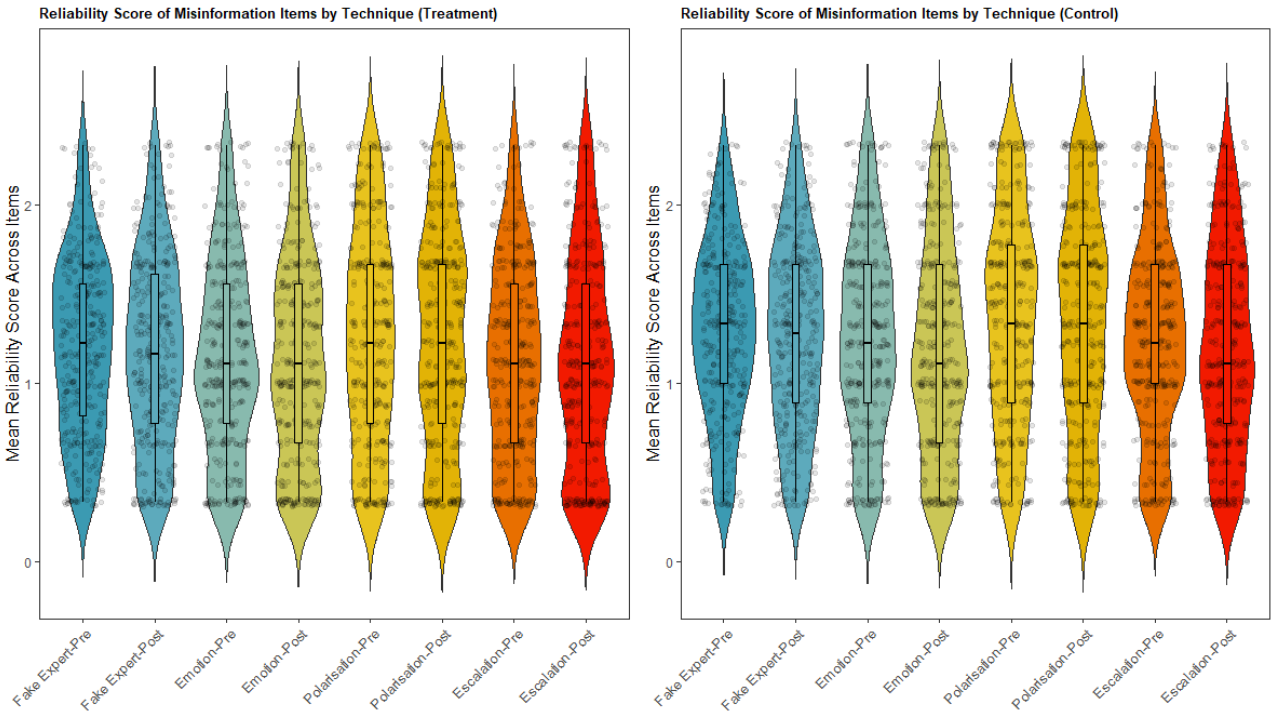


Figure 5 Distribution of mean reliability scores of misinformation items by manipulation technique.

game-based intervention was conducted on participants with minimal experience with operating digital devices. This is compounded by the fact that the majority of our sample was female (55%), who typically have lower digital literacy in this area (Rowntree et al., **2020**). This could have hindered the intervention's efficacy. Furthermore, data quality was poor: only 26% of individuals who played the inoculation game put in the password correctly. Further analysis, however, demonstrated that this did not make a difference to our results (please see Tables S36-38).

Secondly, the game departed from previous game-based inoculation experiments in that it changed the player's perspective from troll to detective. Although this change preserved the critical element of 'active' inoculation that has been effective previously (Pfau et al., **2005**; Roozenbeek & van der Linden, **2019**), it is possible that the role of being not only a detective, but also being undercover, added further layers of complexity that minimized goal salience and clarity for participants, thus reducing its effectiveness. Practitioners may also consider running naturalistic studies in developing countries by conducting interventions broadcasted on WhatsApp through local organizations' subscription lists for increased data availability (Bowles et al., **2020**), or even by artificially constructing a social network in the lab (Pogorelskiy & Shum, **2017**).

Our study may be taken as a lesson in conducting interventions in underexplored populations. In particular, the typical data quality, representativeness, and methodological best practices for running such online experiments in India, and non-WEIRD countries in general, is poorly understood and can impede the experimental process. Campbell-Smith and Bradshaw (**2019**) notes, "having digital connectivity does not mean people are digitally equipped to use online surveys. They have issues in reading and writing, but not in talking." Although we partnered with a local NGO in India, one must also account for gaps in the implementation of scientific experimental designs in the field, particularly by non-academic partners as

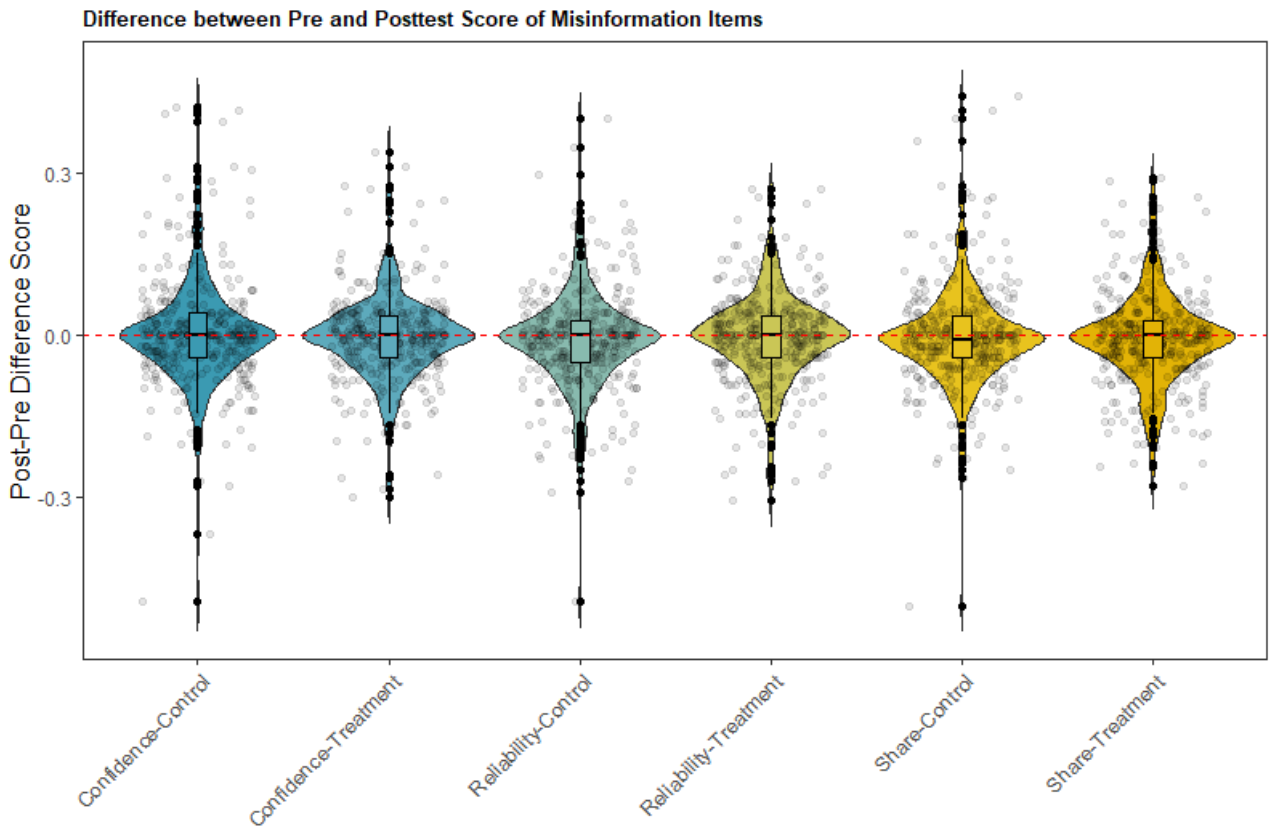


Figure 6 Distribution of post-pre differences between control and treatment groups. Red line drawn at y = 0.

Original Purpose

This paper aims to address the paucity of empirical research investigating misinformation interventions in developing countries. One important difference in the circulation of misinformation in developing countries is its spread through private, encrypted networks such as WhatsApp, which poses different challenges than (the circulation of) misinformation on open networks such as Twitter and Facebook. As such, this paper features a study testing the efficacy of an "inoculation" game in India. We hypothesized that previously reported effects of this inoculation game would be replicated by reducing the reported reliability and sharing intent of misinformation while increasing people's confidence in their own assessments.

it can increase the possibility of unobserved extraneous variables. Additionally, we observed non-random missingness in the data. We find that being assigned to the treatment group increases the odds of an incomplete or missing response, which may have introduced a bias in the results. However, as we found null results no further correction analysis was conducted. Future replications, particularly that find significant results, should pay attention to any differential attrition.

Future studies may also benefit from stronger local relationships (Sircar & Chauchard, **2019**) as well as a greater accountability of the diversity within countries, such as India, that have notable heterogeneity beyond age, gender, and education level (Deshmukh, **2019**). For example, the question on political ideology in this study was more accurately asking people how "free" their

ideology is rather than measuring their political ideology on a left-right scale (measure detailed in the supplement). Although India has been historically classified as clientelist and thus there is no established scale to capture political ideology, some evidence suggests voting behavior among certain groups is not clientelist (Chibber & Verma, **2018**). Future research will need to account for this in the design of surveys. In the context of misinformation, educational interventions have shown differing efficacy depending on political party support (Badrinathan, **2021**) while polarizing content on the basis of religion and caste is often featured in misinformation circulated in India (Al-Zaman, **2021**; Arun, **2019**; Campbell-Smith & Bradshaw, **2019**). For digital interventions, Indian samples may also vary in levels of digital literacy by caste and consumption levels (Mothkoor & Mumtaz, **2021**). Therefore, additional measures, such as whether someone is part of a scheduled group (caste or tribe), religion, income level, and political party affiliation can facilitate a richer understanding of the intervention efficacy in subgroups due to heterogeneity in local factors. To isolate the effect of culture, experiments may also aspire to reach a more digitally literate population within non-WEIRD cultures, given that middle class, urban population in non-WEIRD countries are more likely to resemble the typically studied WEIRD population (Ghai, **2021**).

Conclusion

This study was motivated by scarcity of studies examining non-WEIRD populations in general (Henrich et al., **2010**), and by the lack of research testing the effectiveness of misinformation interventions in democracies such as India (Badrinathan, **2021**), that are being threatened by the prevalence of misinformation. We find null results of a game-based inoculation intervention, *Join this Group,* on ratings of reliability, reported intent to share, and confidence in judgments of misinformation messages. Previous similar game-based inoculation interventions have been demonstrably successful (Basol et al., **2020**; Roozenbeek & van der Linden, **2018**, **2019**, **2020**). We would thus conclude that the results reported here are more likely to reflect an interplay of cultural and experimental design factors. Taken together, we interpret these findings as a call for further adaptation and testing of inoculation interventions on non-WEIRD populations. Modifications may include measuring conceptual mediators such as motivational threat to elucidate and hypothesize potential differences in cross-cultural mechanisms, partnering with local researchers and universities, measuring digital literacy, as well as assessing of behavioral outcomes such as news sharing online.

Acknowledgments

We would like to thank our partners Digital Empowerment Foundation in India for implementing the survey and WhatsApp/Meta for funding.

Funding

This research was funded by WhatsApp through their Research Awards for Social Science and Misinformation program.

References

Adewole, K. S., Anuar, N. B., Kamsin, A., Varathan, K. D., & Razak, S. A. (2017). Malicious accounts: Dark of the social networks. *Journal of Network and Computer Applications*, *79*, 41–67. https://doi.org/10.1016/j.jnca.2016.11.030 (see p. 6).

Al-Zaman, M. S. (2021). A thematic analysis of misinformation in India during the COVID-19 pandemic. *International Information and Library Review*, 1–11. https://doi.org/10.1080/10572317.2021.1908063 (see pp. 1, 13).

Arun, C. (2019). On WhatsApp, rumours, lynchings, and the Indian government. *Economic & Political Weekly*, *54*(6), 30–35. https://papers.ssrn.com/sol3/papers.cfm?abstract%5C_id=3336127 (see pp. 1, 13).

Badrinathan, S. (2021). Educative interventions to combat misinformation: Evidence from a field experiment in india. *American Political Science Review*, *115*(4), 1325–1341. https://doi.org/10.1017/S0003055421000459 (see pp. 1, 2, 9, 13).

Banaji, S., Bhat, R., Agarwal, A., Passanha, N., & Sadhana Pravin, M. (2019). WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India. http://eprints.lse.ac.uk/104316/ (see pp. 1, 2).

Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, *39*(2), 184–207. https://doi.org/10.1111/hcre.12000 (see p. 3).

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281–311. https://doi.org/10.1080/03637751003758193 (see p. 3).

Banas, J. A., & Richards, A. S. (2017). Apprehension or motivation to defend attitudes? exploring the underlying threat mechanism in inoculation-induced resistance to persuasion. *Communication Monographs*, *84*(2), 164–178. https://doi.org/10.1080/03637751.2017.1307999 (see p. 10).

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, *8*(1), 205395172110138. https://doi.org/10.1177/20539517211013868 (see pp. 4, 5, 9).

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1), 2. https://doi.org/10.5334/joc.91 (see pp. 3, 5, 13).

Basol, M., Roozenbeek, J., & van der Linden, S. (2022). Gamified inoculation against misinformation on WhatsApp (see p. 9).

BBC Monitoring. (2021, May 16). *Israel-palestinian conflict: False and misleading claims fact-checked*. https://www.bbc.co.uk/news/57111293 (see pp. 4, 6).

Bhattacharjee, B., Pansari, S., & Dutta, A. (2021). Internet adoption in India. https://images.assettype.com/afaqs/2021-06/b9a3220f-ae2f-43db-a0b4-36a372b243c4/KANTAR%5C_ICUBE%5C_2020%5C_Report%5C_C1.pdf (see p. 2).

Bonetto, E., Troïan, J., Varet, F., Monaco, G., & Girandola, F. (2018). Priming resistance to persuasion decreases adherence to conspiracy theories. *Social Influence*, *13*(3), 125–136. https://doi.org/10.1080/15534510.2018.1471415 (see p. 3).

Bowles, J., Larreguy, H., & Liu, S. (2020). Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in zimbabwe. *PLoS ONE*, *15*(10), 0240005. https://doi.org/10.1371/journal.pone.0240005 (see pp. 2, 11).

Campbell-Smith, U., & Bradshaw, S. (2019). *Global cyber troops country profile: India*. Oxford Internet Institute, University of Oxford. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/India-Profile.pdf (see pp. 2, 11, 13).

Chibber, K. P., & Verma, R. (2018). The myth of cote buying in India. *Ideology and Identity: The Changing Party Systems of India*, 103–130 (see pp. 1, 13).

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0 (see p. 2).

Compton, J. (2012). Inoculation theory. In *The sage handbook of persuasion: Developments in theory and practice* (pp. 220–236). Sage Publications, Inc. (See p. 3).

Compton, J. (2021). Threat and/in inoculation theory. *International Journal of Communication (Online*, *15*(13), 4294–4307. https://ijoc.org/index.php/ijoc/article/view/17634 (see p. 10).

Compton, J., Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6). https://doi.org/10.1111/spc3.12602 (see p. 3).

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE*, *12*(5), 0175799. https://doi.org/10.1371/journal.pone.0175799 (see p. 3).

Deshmukh, Y. (2019). Methodological issues and problems of conducting surveys in India. a commentary by the Indian issp partner organization. *International Journal of Sociology*, *49*(5-6), 400–411. https://doi.org/10.1080/00207659.2019.1683286 (see p. 12).

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781 (see p. 8).

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y (see p. 2).

Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by twitter bots? *First Monday*, *25*(6). https://doi.org/10.5210/fm.v25i6.10633 (see p. 1).

Findlay, S. (2019, February 6). *WhatsApp says Indian rules on encryption 'not possible' to meet*. https://www.ft.com/content/9fcfa604-2a0d-11e9-88a4-c32129756dd8 (see p. 4).

Ghai, S. (2021). It's time to reimagine sample diversity and retire the weird dichotomy. *Nature Human Behaviour*, *5*(8), 971–972. https://doi.org/10.1038/s41562-021-01175-9 (see p. 13).

Goga, O., Venkatadri, G., & Gummadi, K. P. (2015–October 30). The doppelgänger bot attack: Exploring identity impersonation in online social networks. *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 141–153. https://doi.org/10.1145/2815675.2815699 (see pp. 4, 6).

Government of India. (2016). *Rural and urban composition of population – census 2011 and 2011*. https://data.gov.in/resource/rural-and-urban-composition-population-census-2001-and-2011 (see p. 10).

Groenendyk, E. (2018). Competing motives in a polarized electorate: Political responsiveness, identity defensiveness, and the rise of partisan antipathy. *Political Psychology*, *39*(S1), 159–171. https://doi.org/10.1111/pops.12481 (see pp. 4, 6).

Gross, K., & Ambrosio, L. D. (2004). Framing emotional response. *Political Psychology*, *25*(1), 1–29. https://www.jstor.org/stable/3792521 (see pp. 4, 6).

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 15536–15545. https://doi.org/10.1073/pnas.1920498117 (see pp. 2, 9).

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X (see pp. 2, 9, 13).

Ivanov, B., Rains, S. A., Dillingham, L. L., Parker, K. A., Geegan, S. A., & Barbati, J. L. (2022). The role of threat and counterarguing in therapeutic inoculation. *Southern Communication Journal*, *87*(1), 15–27. https://doi.org/10.1080/1041794X.2021.1983012 (see p. 10).

Iyengar, A., Gupta, P., & Priya, N. (2022). Inoculation against conspiracy theories: A consumer side approach to India's fake news problem. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.3995 (see p. 9).

Iyengar, S., & Krupenkin, M. (2018). The strengthening of partisan affect. *Political Psychology*, *39*, 201–218. https://doi.org/10.1111/pops.12487 (see pp. 4, 6).

Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, *47*(8), 459–469. https://doi.org/10.1111/jasp.12453 (see p. 3).

Jung, A. M. (2011). Twittering away the right of publicity: Personality rights and twittering away the right of publicity: Personality rights and celebrity impersonation on social networking websites. *Symposium on Energy Law Article*, *86*(1). https://scholarship.kentlaw.iit.edu/cklawreview/vol86/iss1/16 (see pp. 4, 6).

Kapoor, S., Hughes, P. C., Baldwin, J. R., & Blue, J. (2003). The relationship of individualism–collectivism and self-construals to communication styles in India and the United States. *International Journal of Intercultural Relations*, *27*(6), 683–700. https://doi.org/10.1016/j.ijintrel.2003.08.002 (see p. 2).

Konijn, E. A. (2012). The role of emotion in media use and effects. In *The oxford handbook of media psychology* (pp. 186–211). Oxford University Press. (See pp. 4, 6).

Kwan, L. Y.-Y. (2016). Anger and perception of unfairness and harm: Cultural differences in normative processes that justify sanction assignment. *Asian Journal of Social Psychology*, *19*(1), 6–15. https://doi.org/10.1111/ajsp.12119 (see p. 10).

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963 (see p. 8).

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Era. Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008 (see p. 1).

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018 (see p. 2).

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983 (see p. 3).

Lim, D. H. (2004). Cross cultural differences in online learning motivation. *Educational Media Interna-*

tional, *41*(2), 163–175. https://doi.org/10.1080/09523980410001685784 (see p. 10).

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315 (see p. 3).

Matsumoto, D., Yoo, S. H., & Nakagawa, S. (2008). Culture, emotion regulation, and adjustment. *Journal of Personality and Social Psychology*, *94*(6), 925–937. https://doi.org/10.1037/0022-3514.94.6.925 (see p. 10).

McGuire, W. J. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *The Journal of Abnormal and Social Psychology*, *63*(2), 326–332. https://doi.org/10.1037/h0048344 (see p. 3).

McGuire, W. J. (1964). Some contemporary approaches. *Advances in Experimental Social Psychology*, *1*, 191–229. https://doi.org/10.1016/S0065-2601(08)60052-0 (see pp. 3, 9).

McLaughlin, T. (2018, December 12). *How WhatsApp fuels fake news and violence in India*. https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/ (see p. 4).

Medeiros, B., & Singh, P. (2021). Addressing misinformation on WhatsApp in India through intermediary liability policy, platform design modification, and media literacy. *Journal of Information Policy*, *10*, 276–298. https://doi.org/10.5325/JINFOPOLI.10.2020.0276 (see p. 9).

Melki, M., & Pickering, A. (2014). Ideological polarization and the media. *Economics Letters*, *125*(1), 36–39. https://doi.org/10.1016/j.econlet.2014.08.008 (see p. 6).

Mesquita, B., & Walker, R. (2003). Cultural differences in emotions: A context for interpreting emotional experiences. *Behaviour Research and Therapy*, *41*(7), 777–793. https://doi.org/10.1016/S0005-7967(02)00189-4 (see p. 10).

Mothkoor, V., & Mumtaz, F. (2021, March 23). *The digital dream: Upskilling India for the future*. https://www.ideasforindia.in/topics/governance/the-digital-dream-upskilling-india-for-the-future.html%5C#:%5Ctextasciitilde%7B%7D:text=Digital%5C%20literacy%5C%20levels%5C%20in%5C%20India%5C&text=Based%5C%20on%5C%20the%5C%20above%5C%20definition,just%5C%2025%5C%25%5C%20in%5C%20rural%5C%20areas. (see pp. 10, 13).

Mumo, M. (2021, August 25). *Protecting burundi's vulnerable media. project syndicate*. https://www.project-syndicate.org/commentary/protecting-press-freedom-in-burundi-by-muthoki-mumo-2021-08 (see p. 1).

Park, A., Bryson, C., Clery, E., Curtice, J., & Philips, M. (2013). *British social attitudes: The 30th report*. NatCen Social Research. https://www.bsa.natcen.ac.uk/latest-report/british-social-attitudes-30/key-findings/introduction.aspx (see p. 5).

Pasquetto, I., Center, S., School, H. K., Jahani, E., Baranovsky, A., & Baum, M. A. (2020). Understanding misinformation on mobile instant messengers (mims) in developing countries. https://shorensteincenter.org/misinformation-on-mims/ (see pp. 1, 2, 4).

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465 (see p. 2).

Pfau, M., Ivanov, B., Houston, B., Haigh, M., Sims, J., Gilchrist, E., Russell, J., Wigley, S., Eckstein, J., & Richert, N. (2005). Inoculation and mental processing: The instrumental role of associative networks in the process of resistance to counterattitudinal influence. *Communication Monographs*, *72*(4), 414–441. https://doi.org/10.1080/03637750500322578 (see pp. 4, 11).

Pogorelskiy, K., & Shum, M. (2017). News sharing and voting on social networks: An experimental study. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2972231 (see p. 11).

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in argentina, nigeria, south africa, and the united kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), 2104235118. https://doi.org/10.1073/pnas.2104235118 (see p. 2).

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115 (see pp. 2, 9).

Rao, M. A., Berry, R., Gonsalves, A., Hastak, Y., Shah, M., & Roeser, R. W. (2013). Globalization and the identity remix among urban adolescents in India. *Journal of Research on Adolescence*, *23*(1), 9–24. https://doi.org/10.1111/jora.12002 (see p. 2).

Rathje, S., Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. proceedings of the national academy of

sciences, 118(26, 2024292118. https://doi.org/10.1073/pnas.2024292118 (see p. 6).

Reis, J. C. S., Melo, P., Garimella, K., & Benevenuto, F. (2020). Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation? *Harvard Kennedy School Misinformation Review*, *1*(5). https://doi.org/10.37016/mr-2020-035 (see pp. 1, 2).

Reznik, M. (2013). Identity theft on social networking sites: Developing issues of internet impersonation. *Touro Law Review*, *29*(2), 455–483. https://digitalcommons.tourolaw.edu/lawreviewAvailable at:https://digitalcommons.tourolaw.edu/lawreview/vol29/iss2/12https://digitalcommons.tourolaw.edu/lawreview/vol29/iss2/12 (see pp. 4, 6).

Robb, A. (2021). Anatomy of a fake news scandal. rolling stone. https://www.rollingstone.com/feature/ anatomy-of-a-fake-news-scandal-125877/ (see pp. 4, 6).

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). Disentangling item and testing effects in inoculation research on online misinformation. *Solomon Revisited. Educational and Psychological Measurement*, *81*(2), 340–362. https:/ / doi.org/ 10.1177/ 0013164420940378 (see pp. 3, 4).

Roozenbeek, J., & van der Linden, S. (2018). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570–580. https://doi.org/10.1080/13669877.2018.1443491 (see pp. 3, 13).

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 65. https://doi.org/10.1057/s41599-019-0279-9 (see pp. 2, 3, 8, 11, 13).

Roozenbeek, J., & van der Linden, S. (2020). Breaking harmony square: A game that "inoculates" against political misinformation. *The Harvard Kennedy School (HKS) Misinformation Review*, *1*(8). https://doi.org/ 10.37016/ mr-2020-4 (see pp. 4, 5, 9, 10, 13).

Rowntree, O., Shahnan, M., Bahia, K., Butler, C., Lindsey, D., & Sibthorpe, C. (2020). The mobile gender gap report 2020. https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2020/05/GSMA-The-Mobile-Gender-Gap-Report-2020.pdf (see p. 11).

Sagarin, B. J., Cialdini, R. B., Rice, W. E., & Serna, S. B. (2002). Dispelling the illusion of invulnerability: The motivations and mechanisms of resistance to persuasion. *Journal of Personality and Social Psychology*, *83*(3), 526–541. https://doi.org/10.1037/0022-3514.83.3.526 (see p. 3).

Shahid, F., Mare, S., & Vashistha, A. (2022). Examining source effects perceptions of fake news in India. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSW1, 89), 1–29. https://doi.org/10.1145/3512936 (see p. 10).

Sircar, N., & Chauchard, S. (2019). Dilemmas and challenges of citizen information campaigns: Lessons from a failed experiment in India. *Information, Accountability, and Cumulative Learning*, 287–312. https://doi.org/10.1017/9781108381390.011 (see p. 12).

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, *26*(6), 301–319. https://doi.org/10.1093/jcmc/zmab010 (see p. 1).

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1948–1961. https://doi.org/10.1037/xlm0000422 (see p. 2).

The Editors of Encyclopaedia Britannica. (2012). Santhal. *Encyclopaedia Britannica*. https://www.britannica.com/topic/Santhal (see p. 10).

Traberg, C., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, *700* (see p. 9).

Vaishnav, M., Jaffrelot, C., Mehta, G., Rej, A., Shrinivasan, R., Sagar, R., & Verma, R. (2019). *The bjp in power: Indian democracy and religious nationalism*. Carnegie Endowment for International Peace. https://carnegieendowment.org/2019/04/04/bjp-in-power-indian-democracy-and-religious-nationalism-pub-78677 (see p. 2).

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *28*(3), 460–467. https://doi.org/10.1038/s41591-022-01713-6 (see pp. 2, 3).

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008. https:/ / doi.org/10.1002/gch2.201600008 (see p. 3).

van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., & Lewandowsky, S. (2017). Inoculating against misinformation (J. Sills, Ed.). *Science*, *358*(6367), 1141–1142. https://doi.org/10.1126/science.aar4533 (see p. 1).

Vasudeva, F., & Barkdull, N. (2020). WhatsApp in India? a case study of social media related lynchings.

Social Identities, *26*(5), 574–589. https://doi-org.ezp.lib.cam.ac.uk/10.1080/13504630.2020.1782730 (see p. 1).

Verma, J., & Triandis, H. C. (2020). The measurement of collectivism in India. *Merging Past, Present, and Future in Cross-Cultural Psychology: Selected papers from the Fourteenth International Congress of the International Association for Cross-Cultural Psychology*, 256–265 (see p. 2).

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894 (see p. 2).

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564 (see p. 2).

Wood, M. L. M. (2007). Rethinking the inoculation analogy: Effects on subjects with differing preexisting attitudes. *Human Communication Research*, *33*(3), 357–378. https://doi.org/10.1111/j.1468-2958.2007.00303.x (see p. 10).

Zollo, F., Novak, P. K., Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., Quattrociocchi, W., & Preis, T. (2015). Emotional dynamics in the age of misinformation. *PLoS ONE*, *10*(9), 0138740. https://doi.org/10.1371/journal.pone.0138740 (see p. 4).

Supplemental Materials

Missing Data

A total of *n* = 1283 consenting individuals began the survey of which *n* = 757 were complete and valid responses used in the analysis. As sample demographics were only collected after the post-test measures, it is not possible to understand the differences in individual characteristics across missing and complete responses. However, after filtering out for the those answered at least one question in the pre-test (*n* = 1038), Little's MCAR test (run in R using the *misty* package) for all three dependent variables (reliability, confidence and sharing) suggested that the data were not missing completely at random, χ^2 (5) = 70.59, $p < 0.001$. Thus, we ran a standard logistic regression (using the *glm* function from the *stats* package in R) to investigate patterns of missing data as a function of pre-test responses. This was done by creating a dummy variable where 1 = missing observation and 0 = complete responses. For the manipulative items, higher pre-test confidence scores slightly reduced the odds of missingness (*OR* = 0.030, [95%*CI*; 0.002,0.431]) and being assigned the treatment group increased the odds of missingness (*OR* = 2.171, [95%*CI*; 1.589, 2.967]). This implies that a higher baseline confidence in assessing the reliability of manipulative items decreases the likelihood of missingness while being assigned to the treatment group increases the likelihood of missingness. All other pre-test measures did not affect the odds of dropout. We were not able to assess whether the missing data was due to demographic factors as these were collected at the end of the study.

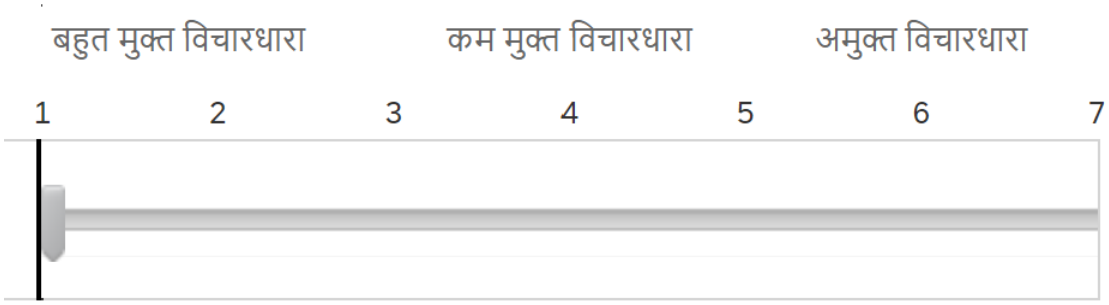
Table S1 Logistic Regression Predicting Missingness (where Missing data = 1, Complete data = 0)

	Odds Ratio	Confidence Intervals
(Intercept)	0.408	*CI* [0.142, 1.172]
Reliability Pre-test (Fake Items)	0.836	*CI* [0.082, 8.490]
Confidence Pre-test (Fake Items)	0.030 **	*CI* [0.002, 0.431]
Sharing Pre-test (Fake Items)	4.965	*CI* [0.509, 48.440]
WhatsApp Usage	1.032	*CI* [0.848, 1.256]
Reliability Pre-test (Real Items)	1.100	*CI* [0.849, 1.425]
Confidence Pre-test (Real Items)	0.811	*CI* [0.626, 1.053]
Sharing Pre-test (Real Items)	1.034	*CI* [0.806, 1.327]
Reliability Pre-test (Real Fake Items)	1.067	*CI* [0.800, 1.424]
Confidence Pre-test (Real Fake Items)	0.874	*CI* [0.660, 1.157]
Sharing Pre-test (Real Fake Items)	1.307	*CI* [0.942, 1.813]
Condition (Treatment)	2.171 ***	*CI* [1.589, 2.967]
N		899
AIC		1057.126
BIC		1114.742
Pseudo R2		0.083

*** p < 0.001; ** p < 0.01; * p < 0.05.

Political Ideology Measurements.

Although we employed a measure from the British Social Attitudes survey, we employed a measure to assess the self-reported identification along the left to right spectrum:



On the slider below, please indicate your political ideology.
[Far left of slider; closer to 1] Very free ideology
[Middle of slider; closer to 4] Less free ideology
[Far right of slider; closer to 7] Not free ideology

Table S2 Sample Composition

Variable	*n*	*Percentage*	*Cumulative Percentage*
Gender			
Male	293	40%	40%
Female	397	55%	95%
Other	35	5%	100%
Age			
18-24	356	49%	49%
25-34	286	39%	89%
35-44	64	9%	97%
45-54	16	2%	100%
55 and over	3	0%	100%
Political Leaning			
1 Very left-wing	139	19%	19%
2	385	53%	72%
3	165	23%	95%
4	34	5%	100%
5 Very right-wing	2	0%	100%
Education			
Class 12	159	22%	22%
Elementary	16	2%	24%
Graduate	306	42%	66%
Post Grad	172	24%	90%
Up to Tenth	72	10%	100%
State			
Bihar	19	3%	3%
Chhattisgarh	42	6%	8%
Delhi	3	0%	9%
Haryana	5	1%	10%
Jharkhand	26	4%	13%
Madhya Pradesh	471	65%	78%
Rajasthan	120	17%	95%
Unknown	6	1%	95%
Uttar Pradesh	33	5%	100%
Frequency of Checking the News			
1 Never	5	1%	1%
2 Occasionally	90	12%	13%
3 Somewhat	166	23%	36%
4 Often	295	41%	77%
5 All the time	169	23%	100%

Table S2 Table S2 continued

Use of social media			
1 Never	28	4%	4%
2 Occasionally	129	18%	22%
3 Somewhat	167	23%	45%
4 Often	212	29%	74%
5 All the time	189	26%	100%
Use of WhatsApp			
1 Never	4	1%	1%
2 Occasionally	22	3%	4%
3 Once a week	26	4%	7%
4 Daily	90	12%	20%
5 More than once a day	520	72%	91%
NA	63	9%	100%
Interest in Politics			
1 Not interested at all	50	7%	7%
2	84	12%	18%
3 Slightly interested	289	40%	58%
4	189	26%	84%
5 Very interested	113	16%	100%

Table S3 ANCOVA on Post-Treatment scores of reliability assessments (of manipulative items)

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.25	1	0.25	36.90	$p < 0.001$		
F_Rel_Pre	6.76	1	6.76	1000.84	$p < 0.001$	.58	[.55, .61]
Condition	0.00	1	0.00	0.00	.969	.00	[.00, 1.00]
Error	4.88	722	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S4 ANCOVA on Post-Treatment scores of confidence measure (of manipulative items)

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.59	1	0.59	83.25	$p < 0.001$		
F_Conf_Pre	6.45	1	6.45	908.30	$p < 0.001$	.56	[.52, .59]
Condition	0.01	1	0.01	1.79	.181	.00	[.00, .01]
Error	5.13	722	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S5 ANCOVA on Post-Treatment scores of sharing measure (of manipulative items)

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.35	1	0.35	48.64	$p < 0.001$		
F_Share_Pre	8.42	1	8.42	1155.91	$p < 0.001$	.62	[.58, .64]
Condition	0.01	1	0.01	1.46	.227	.00	[.00, .01]
Error	5.26	722	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S6 *Bayesian paired sample t-test on dependent variables*

Variable			Statistic	Error %
Reliability of Fake Messages				
Reliability-Post	Reliability-Pre	$BF_{10,\ prior\ =\ 0.707}$	0.249	2.604E^{-08}
Confidence in judgement of Fake Messages				
Confidence-Post	Confidence-Pre	$BF_{10,\ prior\ =\ 0.707}$	0.043	1.612E^{-07}
Intent to share Fake Messages				
Share-Post	Share-Pre	$BF_{10,\ prior\ =\ 0.707}$	0.073	9.425E^{-08}

Table S7 Reliability measure - Fixed-Effects ANCOVA on post-test fake expert manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	6.49	1	6.49	48.51	$p < 0.001$		
FE_Rel_Pre	102.89	1	102.89	768.80	$p < 0.001$	.52	[.48, .55]
Condition	0.03	1	0.03	0.21	.648	.00	[.00, .01]
Error	96.62	722	0.13				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S8 Reliability Measure - ANCOVA on post-test score of emotional manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	14.19	1	14.19	69.20	$p < 0.001$		
EM_Rel_Pre	89.90	1	89.90	438.44	$p < 0.001$	.38	[.33, .42]
Condition	0.04	1	0.04	0.21	.649	.00	[.00, .01]
Error	148.04	722	0.21				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S9 Reliability Measure - ANCOVA on post-test score of polarisation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	16.52	1	16.52	81.00	$p < 0.001$		
PL_Rel_Pre	119.46	1	119.46	585.71	$p < 0.001$	.45	[.41, .49]
Condition	0.07	1	0.07	0.35	.553	.00	[.00, .01]
Error	147.26	722	0.20				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S10 Reliability Measure - ANCOVA on post-test score of escalation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	13.38	1	13.38	68.51	$p < 0.001$		
ES_Rel_Pre	96.59	1	96.59	494.40	$p < 0.001$	.41	[.36, .45]
Condition	0.01	1	0.01	0.03	.852	.00	[.00, .00]
Error	141.06	722	0.20				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S11 *Reliability measure – ANCOVA on post-test score of authentic fake news items*

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	63.45	1	63.45	116.01	.000		
RF_Rel_Pre	198.77	1	198.77	363.43	.000	.34	[.29, .38]
Condition	0.54	1	0.54	0.99	.319	.00	[.00, .01]
Error	389.40	712	0.55				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S12 Reliability measure - ANCOVA on post-test score of real (non-manipulative) items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	partial η^2	partial η^2 90%*CI* [LL, UL]
(Intercept)	41.14	1	41.14	81.48	.000		
R_Rel_Pre	265.29	1	265.29	525.46	.000	.42	[.38, .46]
Condition	0.05	1	0.05	0.09	.763	.00	[.00, .00]
Error	364.52	722	0.50				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S13 Confidence measure - ANCOVA on post-test score of fake expert manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	partial η^2	partial η^2 90%*CI* [LL, UL]
(Intercept)	16.56	1	16.56	123.94	$p < 0.001$		
FE_Conf_Pre	96.67	1	96.67	723.45	$p < 0.001$	.50	[.46, .54]
Condition	0.21	1	0.21	1.56	.211	.00	[.00, .01]
Error	96.47	722	0.13				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S14 Confidence measure – ANCOVA on post-test score of emotional manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	partial η^2	partial η^2 90%*CI* [LL, UL]
(Intercept)	19.49	1	19.49	103.28	$p < 0.001$		
EM_Conf_Pre	90.92	1	90.92	481.79	$p < 0.001$	.40	[.36, .44]
Condition	0.20	1	0.20	1.05	.306	.00	[.00, .01]
Error	136.25	722	0.19				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S15 Confidence measure – ANCOVA on post-test score of polarisation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	22.17	1	22.17	119.57	$p < 0.001$		
PL_Conf_Pre	92.32	1	92.32	497.93	$p < 0.001$	.41	[.37, .45]
Condition	0.22	1	0.22	1.18	.278	.00	[.00, .01]
Error	133.87	722	0.19				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S16 Confidence measure – ANCOVA on post-test score of escalation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	24.97	1	24.97	135.70	$p < 0.001$		
ES_Conf_Pre	90.68	1	90.68	492.84	$p < 0.001$	.41	[.36, .45]
Condition	0.21	1	0.21	1.17	.280	.00	[.00, .01]
Error	132.84	722	0.18				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S17 Confidence measure – ANCOVA on post-test score of authentic fake news items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	72.95	1	72.95	143.40	.000		
RF_Conf_Pre	200.25	1	200.25	393.60	.000	.36	[.31, .40]
Condition	0.86	1	0.86	1.68	.195	.00	[.00, .01]
Error	361.73	711	0.51				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S18 Confidence measure – ANCOVA on post-test score of real (non-manipulative) items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	64.69	1	64.69	110.28	.000		
R_Conf_Pre	325.73	1	325.73	555.32	.000	.43	[.39, .47]
Condition	0.65	1	0.65	1.10	.295	.00	[.00, .01]
Error	423.49	722	0.59				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S19 Sharing Measure – ANCOVA on post-test score of fake expert manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	10.08	1	10.08	69.52	$p < 0.001$		
FE_Share_Pre	125.70	1	125.70	867.08	$p < 0.001$	.55	[.51, .58]
Condition	0.28	1	0.28	1.94	.164	.00	[.00, .01]
Error	104.67	722	0.14				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S20 Sharing Measure – ANCOVA on post-test score of emotional manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	13.68	1	13.68	67.67	$p < 0.001$		
EM_Share_Pre	133.09	1	133.09	658.41	$p < 0.001$	.48	[.44, .51]
Condition	0.06	1	0.06	0.29	.590	.00	[.00, .01]
Error	145.95	722	0.20				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S21 Sharing Measure – ANCOVA on post-test score of polarisation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	17.68	1	17.68	82.23	$p < 0.001$		
PL_Share_Pre	130.23	1	130.23	605.59	$p < 0.001$	.46	[.41, .49]
Condition	0.59	1	0.59	2.75	.098	.00	[.00, .01]
Error	155.26	722	0.22				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S22 Sharing Measure – ANCOVA on post-test score of escalation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	17.43	1	17.43	88.22	$p < 0.001$		
ES_Share_Pre	130.11	1	130.11	658.64	$p < 0.001$	.48	[.44, .51]
Condition	0.55	1	0.55	2.77	.097	.00	[.00, .01]
Error	142.63	722	0.20				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S23 Sharing measure – ANCOVA on post-test score of authentic fake news items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	49.16	1	49.16	83.26	.000		
RF_Share_Pre	291.58	1	291.58	493.82	.000	.41	[.37, .45]
Condition	0.07	1	0.07	0.12	.732	.00	[.00, .00]
Error	414.50	702	0.59				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S24 Sharing Measure – ANCOVA on post-test score of real (non-manipulative) items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	44.18	1	44.18	78.11	.000		
R_Share_Pre	373.52	1	373.52	660.31	.000	.48	[.44, .51]
Condition	0.79	1	0.79	1.39	.239	.00	[.00, .01]
Error	408.41	722	0.57				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S25 Pre-Post Mean Differences

Variable	*Condition*	*N*	*Mean.Difference*	*SD*
Reliability (manipulative items)	Treatment	360	-0.00	0.08
Confidence (manipulative items)	Treatment	360	-0.00	0.08
Sharing (manipulative items)	Treatment	360	-0.00	0.08
Reliability (real items)	Treatment	360	0.07	0.73
Confidence (real items)	Treatment	360	-0.02	0.83
Sharing (real items)	Treatment	360	0.05	0.80
Reliability (authentic fake items)	Treatment	355	0.04	0.78
Confidence (authentic fake items)	Treatment	357	-0.01	0.78
Sharing (authentic fake items)	Treatment	352	0.01	0.84
Reliability (manipulative items)	Control	365	-0.01	0.09
Confidence (manipulative items)	Control	365	0.00	0.10
Sharing (manipulative items)	Control	365	-0.00	0.09
Reliability (real items)	Control	365	0.04	0.80
Confidence (real items)	Control	365	0.04	0.84
Sharing (real items)	Control	365	0.07	0.81
Reliability (authentic fake items)	Control	360	0.03	0.85
Confidence (authentic fake items)	Control	357	0.04	0.81
Sharing (authentic fake items)	Control	353	-0.03	0.84

Table S26 Reliability Measure - item-level ANOVA table (pre-post difference scores)

Variable	*F.value*	*df1*	*df2*	*p*
Diff_Fake_Rel_1-FakeExp	0.271	1	723	0.603
Diff_Fake_Rel_2-FakeExp	0.108	1	723	0.743
Diff_Fake_Rel_3-FakeExp	0.303	1	723	0.582
Diff_Fake_Rel_4-Emotion	0.044	1	723	0.834
Diff_Fake_Rel_5-Emotion	3.286	1	723	0.070
Diff_Fake_Rel_6-Polarise	0.371	1	723	0.543
Diff_Fake_Rel_7-Emotion	1.407	1	723	0.236
Diff_Fake_Rel_8-Polarise	1.744	1	723	0.187
Diff_Fake_Rel_9-Polarise	0.010	1	723	0.919
Diff_Fake_Rel_10-Escalate	0.322	1	723	0.571
Diff_Fake_Rel_11-Escalate	1.317	1	723	0.252
Diff_Fake_Rel_12-Escalate	0.008	1	723	0.930
Diff_Real_Fake_Conf_13 (authentic fake item)	0.186	1	713	0.666
Diff_Real_Fake_Conf_14 (authentic fake item)	0.283	1	723	0.595
Diff_Real_Rel_15	0.191	1	723	0.662
Diff_Real_Rel_16	1.345	1	723	0.247

Table S27 Confidence Measure - item-level ANOVA table (pre-post difference scores)

Variable	*F.value*	*df1*	*df2*	*p*
Diff_Fake_Conf_1-FakeExp	2.323	1	723	0.128
Diff_Fake_Conf_2-FakeExp	0.001	1	723	0.974
Diff_Fake_Conf_3-FakeExp	0.000	1	723	0.990
Diff_Fake_Conf_4-Emotion	0.214	1	723	0.644
Diff_Fake_Conf_5-Emotion	0.576	1	723	0.448
Diff_Fake_Conf_6-Polarise	2.496	1	723	0.115
Diff_Fake_Conf_7-Emotion	0.327	1	723	0.567
Diff_Fake_Conf_8-Polarise	1.697	1	723	0.193
Diff_Fake_Conf_9-Polarise	3.400	1	723	0.066
Diff_Fake_Conf_10-Escalate	0.035	1	723	0.851
Diff_Fake_Conf_11-Escalate	0.929	1	723	0.336
Diff_Fake_Conf_12-Escalate	0.807	1	723	0.369
Diff_Real_Fake_Conf_13 (authentic fake item)	2.458	1	712	0.117
Diff_Real_Fake_Conf_14 (authentic fake item)	0.337	1	723	0.562
Diff_Real_Conf_15	2.044	1	723	0.153
Diff_Real_Conf_16	0.002	1	723	0.965

Table S28 Sharing Measure - item-level ANOVA table (pre-post difference scores)

Variable	*F.value*	*df1*	*df2*	*p*
Diff_Fake_Share_1-FakeExp	0.592	1	723	0.442
Diff_Fake_Share_2-FakeExp	0.385	1	723	0.535
Diff_Fake_Share_3-FakeExp	0.004	1	723	0.952
Diff_Fake_Share_4-Emotion	0.012	1	723	0.911
Diff_Fake_Share_5-Emotion	1.179	1	723	0.278
Diff_Fake_Share_6-Polarise	0.233	1	723	0.629
Diff_Fake_Share_7-Emotion	0.010	1	723	0.921
Diff_Fake_Share_8-Polarise	1.426	1	723	0.233
Diff_Fake_Share_9-Polarise	0.672	1	723	0.413
Diff_Fake_Share_10-Escalate	2.935	1	723	0.087
Diff_Fake_Share_11-Escalate	0.146	1	723	0.703
Diff_Fake_Share_12-Escalate	0.144	1	723	0.705
Diff_Real_Fake_Conf_13 (authentic fake item)	0.664	1	703	0.415
Diff_Real_Fake_Conf_14 (authentic fake item)	0.099	1	723	0.753
Diff_Real_Share_15	0.006	1	723	0.936
Diff_Real_Share_16	0.203	1	723	0.653

Table S29 Reliability Measure – Item-level statistics

	Treatment				**Control**			
Item	*Mpre*	*SDpre*	*Mpost*	*SDpost*	*Mpre*	*SDpre*	*Mpost*	*SDpost*
Fake_Rel_1-FakeExp	4.51	2.27	4.28	2.30	4.50	2.27	4.36	2.26
Fake_Rel_10-Escalate	3.69	2.17	3.50	2.18	4.11	2.24	3.83	2.27
Fake_Rel_11-Escalate	2.95	1.98	3.09	2.07	3.24	2.13	3.20	2.08
Fake_Rel_12-Escalate	3.59	2.15	3.55	2.13	3.84	2.15	3.82	2.10
Fake_Rel_2-FakeExp	3.12	2.08	3.19	2.13	3.40	2.16	3.42	2.14
Fake_Rel_3-FakeExp	3.92	2.26	3.77	2.21	4.18	2.29	4.12	2.24
Fake_Rel_4-Emotion	3.44	1.97	3.36	2.05	3.63	2.09	3.59	2.10
Fake_Rel_5-Emotion	3.68	2.13	3.61	2.12	4.15	2.14	3.78	2.09
Fake_Rel_6-Polarise	4.03	2.26	3.87	2.25	4.21	2.20	4.16	2.22
Fake_Rel_7-Emotion	3.24	2.07	3.31	2.16	3.50	2.05	3.38	2.15
Fake_Rel_8-Polarise	3.53	2.06	3.61	2.18	3.95	2.06	3.82	2.14
Fake_Rel_9-Polarise	3.79	2.21	3.82	2.17	4.07	2.18	4.12	2.17
Real_Fake Rel_13	3.26	2.06	3.36	2.06	3.45	2.03	3.65	2.06
Real_Fake Rel_14	3.32	2.18	3.34	2.28	3.76	2.29	3.69	2.28
Real_Rel_15	2.49	1.93	2.59	1.98	2.77	2.10	2.93	2.13
Real_Rel_16	2.65	2.00	2.82	2.11	2.89	2.06	2.89	2.04

Table S30 Confidence Measure – Item-level statistics

	Treatment				Control			
Item	*Mpre*	*SDpre*	*Mpost*	*SDpost*	*Mpre*	*SDpre*	*Mpost*	*SDpost*
Fake_Conf_1-FakeExp	4.80	2.01	4.84	1.94	4.83	2.12	5.09	1.93
Fake_Conf_10-Escalate	4.57	2.02	4.56	1.96	4.68	2.00	4.69	1.98
Fake_Conf_11-Escalate	4.51	2.04	4.50	2.07	4.42	2.19	4.56	2.05
Fake_Conf_12-Escalate	4.60	2.01	4.49	1.93	4.57	2.00	4.60	1.96
Fake_Conf_2-FakeExp	4.32	2.03	4.36	2.04	4.47	2.12	4.50	2.10
Fake_Conf_3-FakeExp	4.66	2.04	4.72	1.94	4.72	2.07	4.79	2.00
Fake_Conf_4-Emotion	4.32	1.99	4.24	1.97	4.42	2.01	4.41	2.03
Fake_Conf_5-Emotion	4.66	1.96	4.54	1.95	4.67	2.04	4.66	1.92
Fake_Conf_6-Polarise	4.85	1.95	4.63	1.98	4.78	2.01	4.81	2.05
Fake_Conf_7-Emotion	4.44	1.98	4.36	2.02	4.39	1.99	4.41	1.97
Fake_Conf_8-Polarise	4.54	1.94	4.64	2.01	4.72	1.95	4.61	1.97
Fake_Conf_9-Polarise	4.71	1.97	4.66	1.99	4.55	2.01	4.79	1.89
Real Fake_Conf_13	4.54	2.02	4.46	2.00	4.55	1.98	4.69	1.87
Real Fake_Conf_14	4.45	2.11	4.50	2.13	4.70	2.14	4.66	2.05
Real_Conf_15	4.41	2.27	4.25	2.25	4.29	2.24	4.35	2.21
Real_Conf_16	4.21	2.26	4.29	2.18	4.40	2.19	4.48	2.20

Table S31 Sharing Measure – Item-level statistics

	Treatment				Control			
Item	*Mpre*	*SDpre*	*Mpost*	*SDpost*	*Mpre*	*SDpre*	*Mpost*	*SDpost*
Fake_Share_1-FakeExp	4.55	2.28	4.39	2.25	4.73	2.26	4.70	2.21
Fake_Share_10-Escalate	4.03	2.24	3.79	2.24	4.32	2.27	4.35	2.21
Fake_Share_11-Escalate	3.37	2.19	3.51	2.21	3.65	2.31	3.73	2.27
Fake_Share_12-Escalate	3.93	2.25	3.83	2.26	4.12	2.17	4.08	2.24
Fake_Share_2-FakeExp	3.57	2.23	3.57	2.18	3.91	2.28	3.82	2.23
Fake_Share_3-FakeExp	3.92	2.26	4.03	2.31	4.30	2.28	4.42	2.25
Fake_Share_4-Emotion	3.58	2.16	3.51	2.19	3.88	2.29	3.83	2.20
Fake_Share_5-Emotion	3.88	2.19	3.80	2.19	4.33	2.21	4.08	2.18
Fake_Share_6-Polarise	4.03	2.31	4.00	2.32	4.55	2.24	4.45	2.23
Fake_Share_7-Emotion	3.38	2.17	3.45	2.23	3.68	2.18	3.77	2.23
Fake_Share_8-Polarise	3.93	2.24	3.78	2.26	4.11	2.18	4.16	2.27
Fake_Share_9-Polarise	4.14	2.22	3.92	2.29	4.29	2.20	4.21	2.14
Real Fake_Share_13	3.62	2.26	3.75	2.23	4.03	2.31	4.04	2.22
Real Fake Share_14	3.84	2.27	3.79	2.29	4.19	2.37	4.09	2.33
Real_Share_15	2.94	2.12	3.05	2.20	3.48	2.28	3.58	2.33
Real_Share_16	3.20	2.30	3.31	2.30	3.40	2.29	3.58	2.34

Table S32 Reliability, Confidence and Sharing Measure of all manipulative items - Two-sided Independent Samples t-test of equivalence (TOSTs)

Var	*b.0.*	*t.0.*	*df.0.*	*p.0.*	*b.1.*	*t.1.*	*df.1.*	*p.1.*	*b.2.*	*t.2.*	*df.2.*	*p.2.*
F_Rel_Post	t-test	1.92	721.68	0.06	TOST Upper	-1.45	721.68	0.07	TOST Lower	5.29	721.68	$p < 0.001$
F_Conf_Post	t-test	1.03	721.43	0.31	TOST Upper	-2.34	721.43	0.01	TOST Lower	4.39	721.43	$p < 0.001$
F_Share_Post	t-test	2.72	719.73	0.01	TOST Upper	-0.64	719.73	0.26	TOST Lower	6.09	719.73	$p < 0.001$

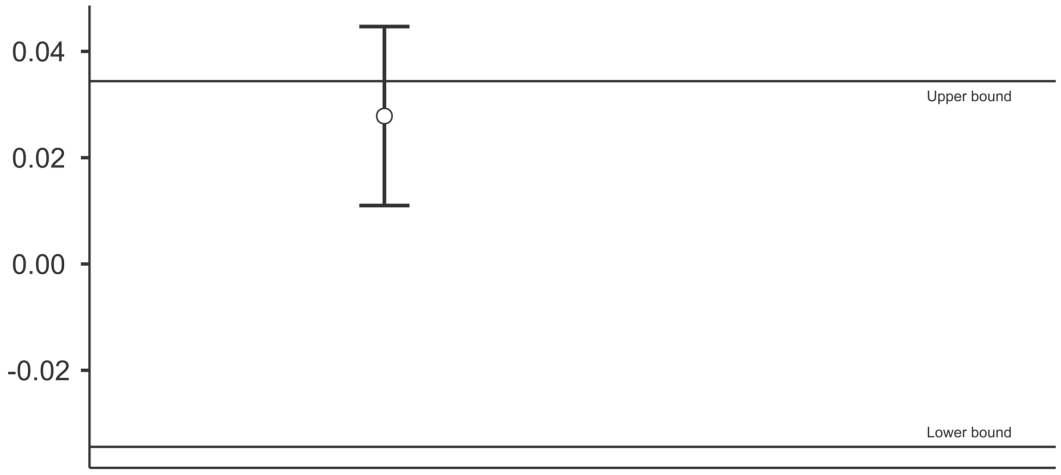


Table S33 Linear regression with difference in pre-post reliability rating of manipulative messaging as the dependent variable

Predictors	*Estimates*	CI	*p*
(Intercept)	−0.02	−0.08 – −0.03	0.436
Condition [Treatment]	0.00	−0.01 – −0.02	0.633
Gender [2]	−0.01	−0.02 – −0.00	0.186
Gender [3]	−0.03	−0.06 – −0.00	0.069
Grad [1]	−0.00	−0.01 – −0.01	0.955
Age25-34	−0.00	−0.02 – −0.01	0.580
Age35-44	0.01	−0.02 – −0.03	0.542
Age45-54	−0.03	−0.08 – −0.01	0.130
Age [55 and over]	0.00	−0.09 – −0.10	0.964
Pol_interest_1	−0.00	−0.01 – −0.01	0.973
LR_Score	−0.00	−0.01 – −0.01	0.906
FromMP [1]	0.01	−0.01 – −0.02	0.517
Lib_Auth	0.01	−0.00 – −0.03	0.054
WAUse_1	−0.01	−0.01 – −0.00	0.198
News.checking_1	0.01	−0.00 – −0.01	0.087
Social.checking_1	−0.00	−0.01 – −0.00	0.527
Observations		662	
R^2 / R^2 adjusted		0.026 / 0.003	

Table S34 Linear regression with difference in pre-post confidence rating of manipulative messaging as the dependent variable

Predictors	*Estimates*	CI %	*p*
(Intercept)	−0.01	−0.07 – −0.05 %	0.783
Condition [Treatment]	−0.01	−0.02 – −0.01 %	0.335
Gender [2]	−0.00	−0.02 – −0.01 %	0.570
Gender [3]	−0.01	−0.04 – −0.02 %	0.622
Grad [1]	−0.01	−0.02 – −0.01 %	0.218
Age25-34	−0.01	−0.02 – −0.01 %	0.492
Age35-44	−0.01	−0.03 – −0.02 %	0.603
Age45-54	−0.04	−0.08 – −0.01 %	0.103
Age [55 and over]	0.00	−0.10 – −0.10 %	0.944
Pol_interest_1	−0.00	−0.01 – −0.00 %	0.421
LR_Score	−0.01	−0.02 – −0.00 %	0.310
FromMP [1]	−0.01	−0.02 – −0.01 %	0.523
Lib_Auth	0.01	−0.01 – −0.02 %	0.253
WAUse_1	0.00	−0.01 – −0.01 %	0.611
News.checking_1	0.01	0.00 – −0.02 %	**0.032**
Social.checking_1	−0.00	−0.01 – −0.00 %	0.203
Observations		662	
R^2 / R^2 adjusted		0.019 / -0.003	

Table S35 Linear regression with difference in pre-post sharing rating of manipulative messaging as the dependent variable

Predictors	*Estimates*	CI	*p*
(Intercept)	0.04	−0.02 – −0.10	0.192
Condition [Treatment]	−0.01	−0.02 – −0.01	0.382
Gender [2]	−0.00	−0.02 – −0.01	0.719
Gender [3]	0.00	−0.03 – −0.03	0.966
Grad [1]	−0.00	−0.02 – −0.01	0.679
Age25-34	0.00	−0.01 – −0.02	0.603
Age35-44	0.00	−0.02 – −0.03	0.779
Age45-54	−0.01	−0.06 – −0.03	0.628
Age [55 and over]	−0.02	−0.12 – −0.09	0.746
Pol_interest_1	−0.01	−0.01 – −0.00	0.059
LR_Score	−0.00	−0.01 – −0.01	0.434
FromMP [1]	−0.00	−0.02 – −0.02	0.935
Lib_Auth	−0.00	−0.02 – −0.01	0.743
WAUse_1	−0.00	−0.01 – −0.00	0.314
News.checking_1	0.01	−0.00 – −0.02	0.102
Social.checking_1	−0.00	−0.01 – −0.00	0.555
Observations		662	
R^2 / R^2 adjusted		0.012 / -0.011	

Table S36 ANCOVA on Post-Treatment scores of reliability assessments (of manipulative items) – data filtered for treatment participants that inputted the game password correctly

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.17	1	0.17	23.97	$p < 0.001$		
F_Rel_Pre	4.18	1	4.18	586.29	$p < 0.001$	.56	[.52, .60]
Condition	0.00	1	0.00	0.10	.752	.00	[.00, .01]
Error	3.25	456	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S37 ANCOVA on Post-Treatment scores of confidence in assessments (of manipulative items) – data filtered for treatment participants that inputted the game password correctly

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.55	1	0.55	70.67	$p < 0.001$		
F_Conf_Pre	3.54	1	3.54	458.50	$p < 0.001$	.50	[.45, .54]
Condition	0.00	1	0.00	0.25	.615	.00	[.00, .01]
Error	3.52	456	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S38 ANCOVA on Post-Treatment scores of sharing measure (of manipulative items) – data filtered for treatment participants that inputted the game password correctly

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.32	1	0.32	42.52	$p < 0.001$		
F_Share_Pre	4.85	1	4.85	652.45	$p < 0.001$	.59	[.54, .63]
Condition	0.00	1	0.00	0.67	.413	.00	[.00, .01]
Error	3.39	456	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S39 Proportion of rural population across participants' states

State	*n*	*Rural population (%)*	*Weighted Rural*[1]
Bihar	19	89	1685
Chhattisgarh	42	77	3226
Delhi	3	2	8
Haryana	5	65	325
Jharkhand	26	76	1976
Madhya Pradesh	471	72	34100
Rajasthan	120	75	9012
Unknown*	6	69	413
Uttar Pradesh	33	78	2564
		Weighted Mean	73.5

**For missing values, rural proportion of India's national population was imputed*
*[1] Weighted Rural = n * Rural population (%)*
All rural population (%) values sourced from:

Table S40 Distribution between conditions by state

Condition	*State*	*n*
Control	Bihar	15
Treatment	Bihar	4
Control	Chhattisgarh	22
Treatment	Chhattisgarh	20
Control	Delhi	1
Treatment	Delhi	2
Control	Haryana	5
Control	Jharkhand	10
Treatment	Jharkhand	16
Control	Madhya Pradesh	232
Treatment	Madhya Pradesh	239
Control	Rajasthan	62
Treatment	Rajasthan	58
Control	Unknown	4
Treatment	Unknown	2
Control	Uttar Pradesh	14
Treatment	Uttar Pradesh	19

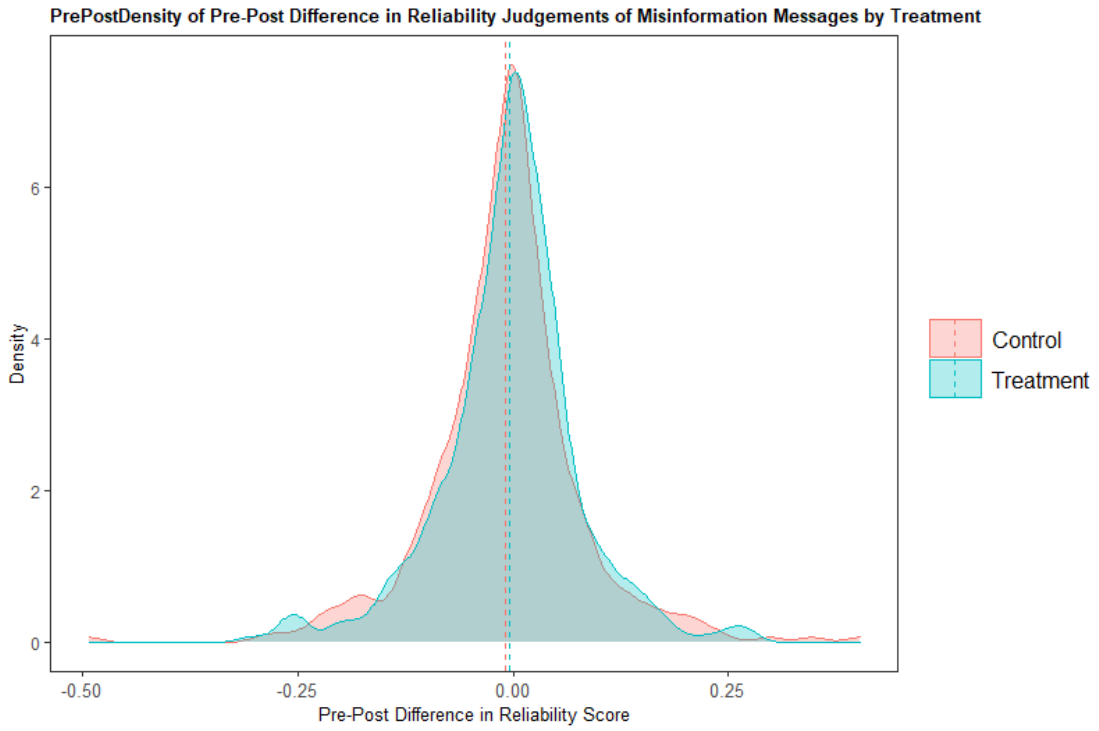


Figure S1 Distribution of Pre-Post Differences in Reliability Judgements of Manipulative Items by Condition

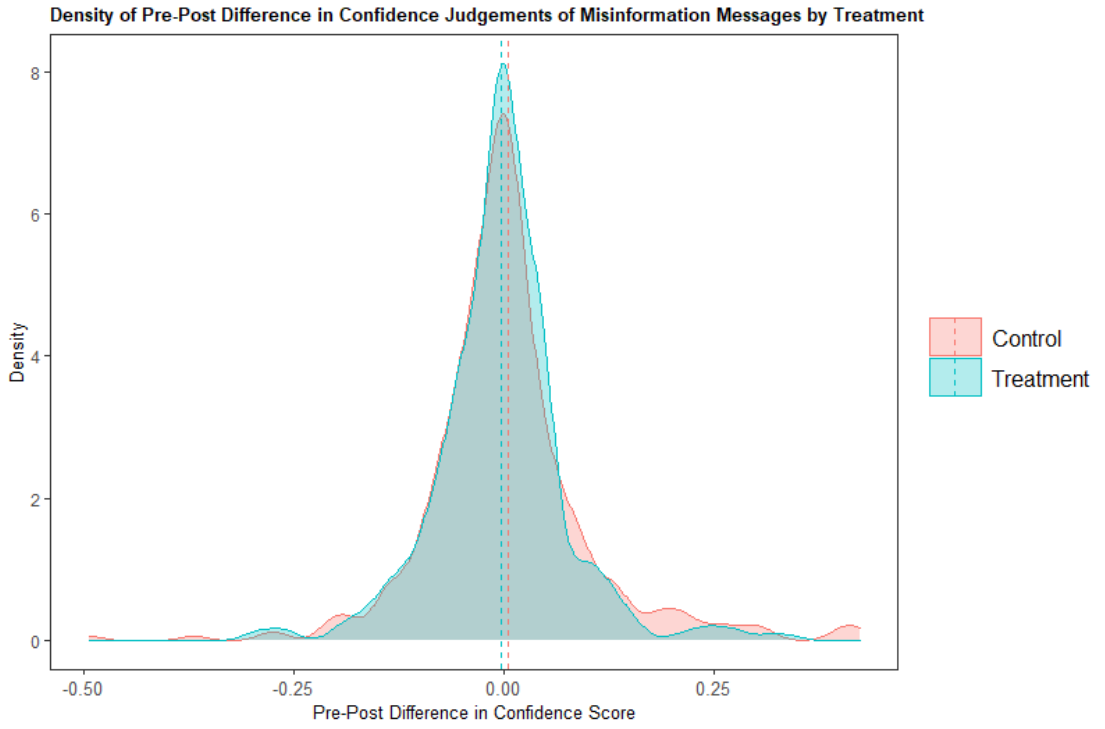


Figure S2 Distribution of Pre-Post Differences in Confidence in Judgements of Manipulative Items by Condition

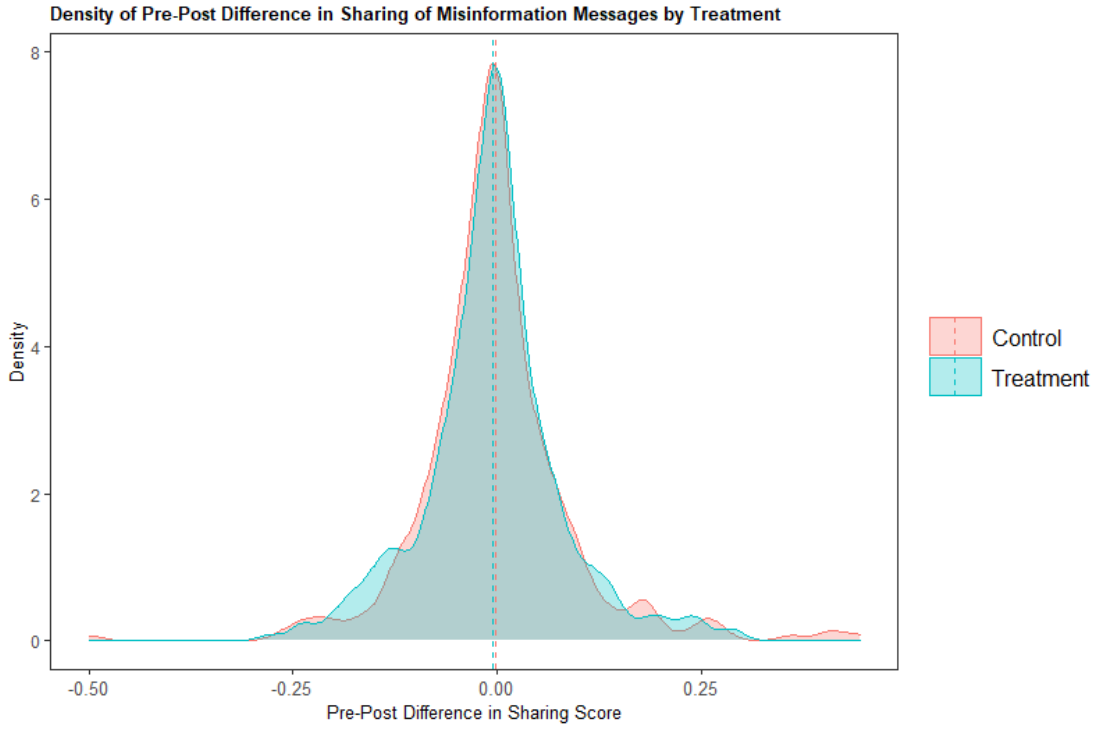


Figure S3 Distribution of Pre-Post Differences in Likelihood to Share Manipulative Items by Condition

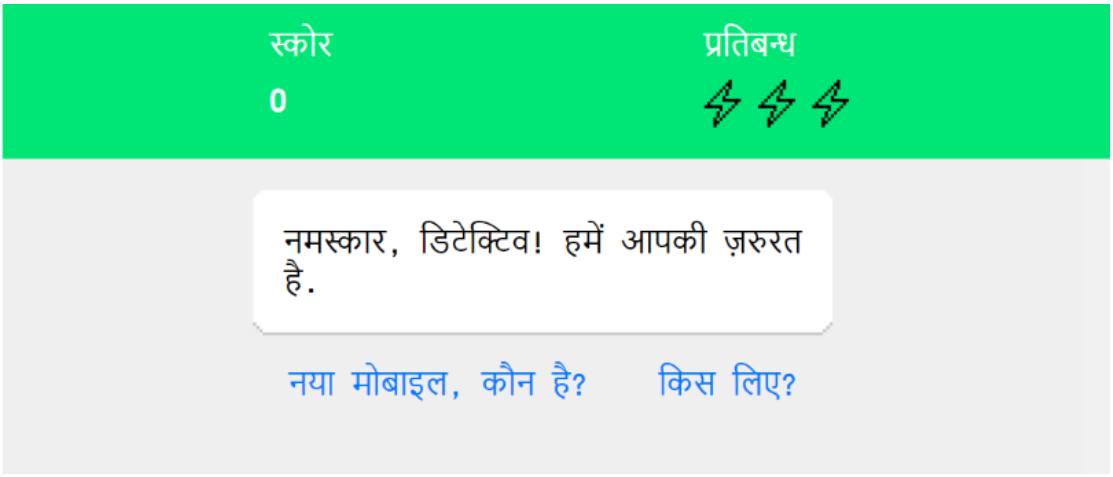


Figure S4 In-game Screenshot - First screen shown after starting the game, introducing the character and motive
Translation:
Green Bar (Left to Right): "Score" "Sanctions"
White Box: "Hello, Detective! We need you"

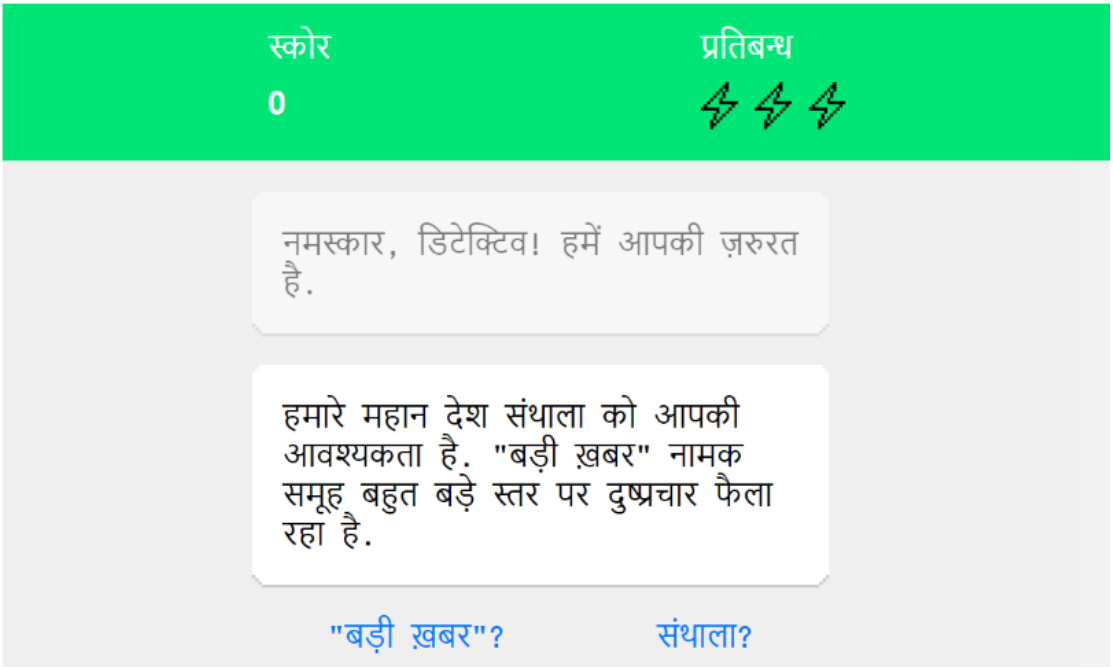


Figure S5 In-game Screenshot - Second screen shown after starting the game, depicting an explanation of the propaganda spreading on WhatsApp.
Translation:
Green Bar (Left to Right): "Score" "Sanctions"
White Box: "Our great country Santhala needs you. A group called "Big News" is spreading propaganda at a very large scale"
Blue Text (Left to Right): "New mobile, who's this?" "For what?"
Blue text: "Big News?" "Santhala?"

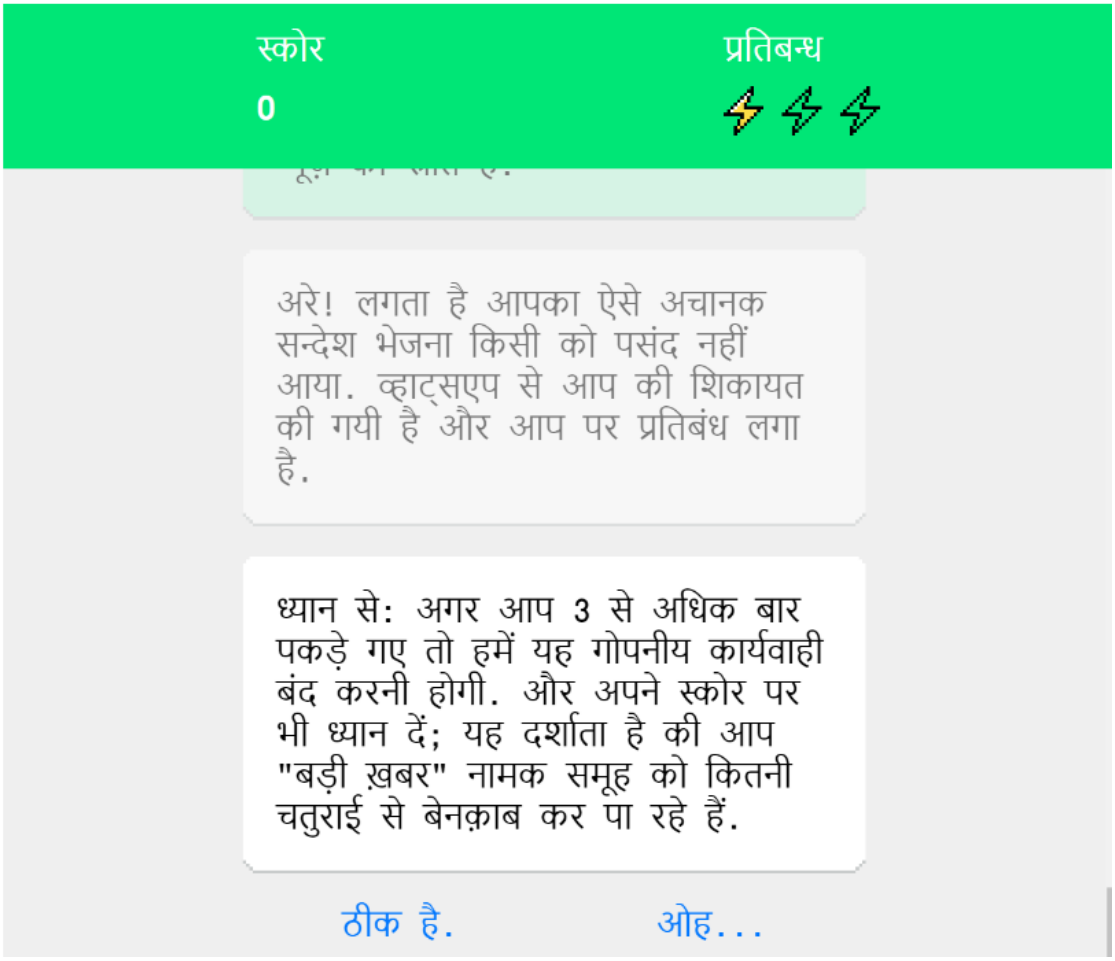


Figure S6 In-game Screenshot - An in-game screenshot explaining the rules of the game.
Translation:
Green Bar (Left to Right): "Score" "Sanctions"
White Box: "Be careful: If you get caught more than 3 times then we have to stop this secrecy. And watch your score as well; this will tell you how much you are exposing the "Bad News" group."
Blue text: "That's fine""Okay..."

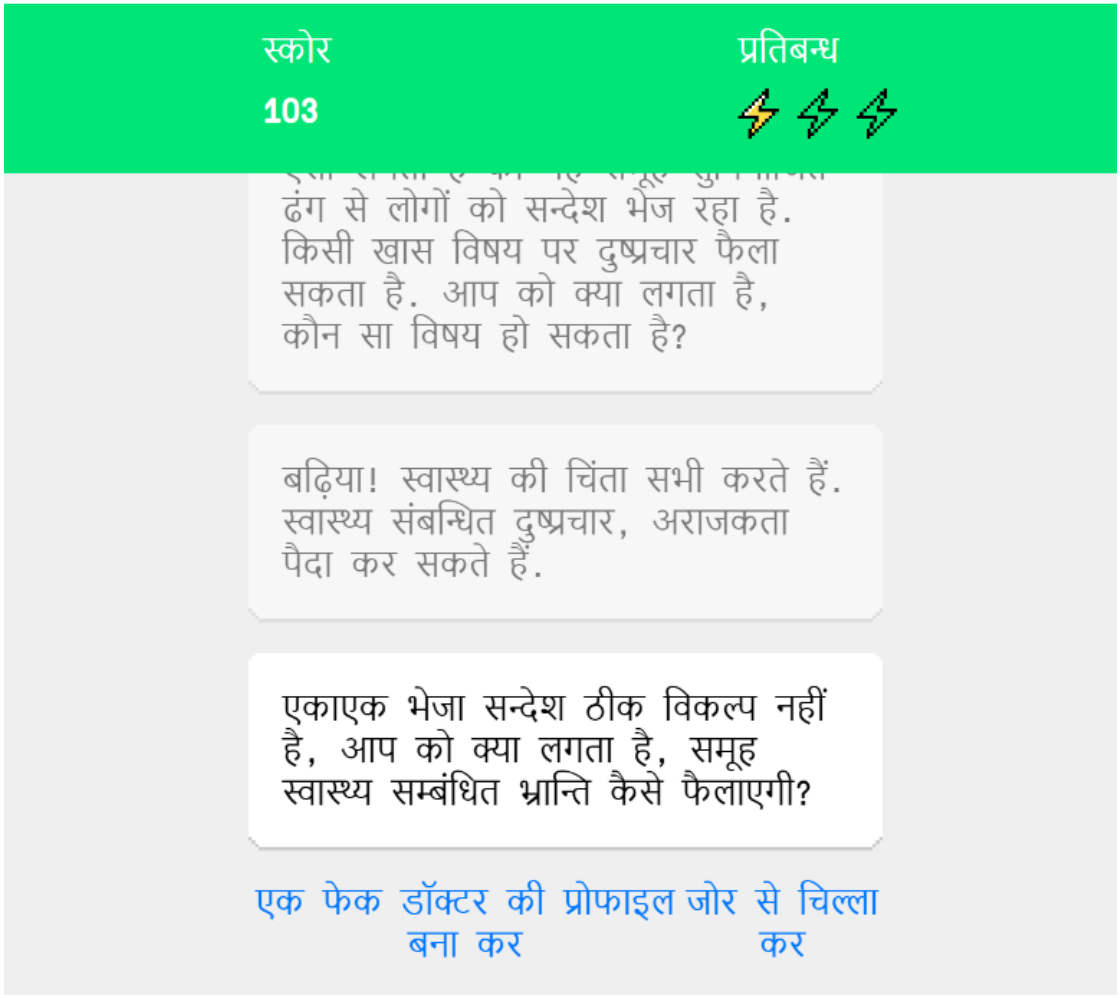


Figure S7 In-game screenshot - Showing how a Fake News technique (using a fake expert) is taught.
Translation: *Green Bar (Left to Right): "Score" "Sanctions"*
White Box: "Just sending a message all of a sudden isn't the right way, what do you think, how will the group spread this health-related misconception?"
Blue text: "By creating a fake doctor profile" "By shouting loudly"

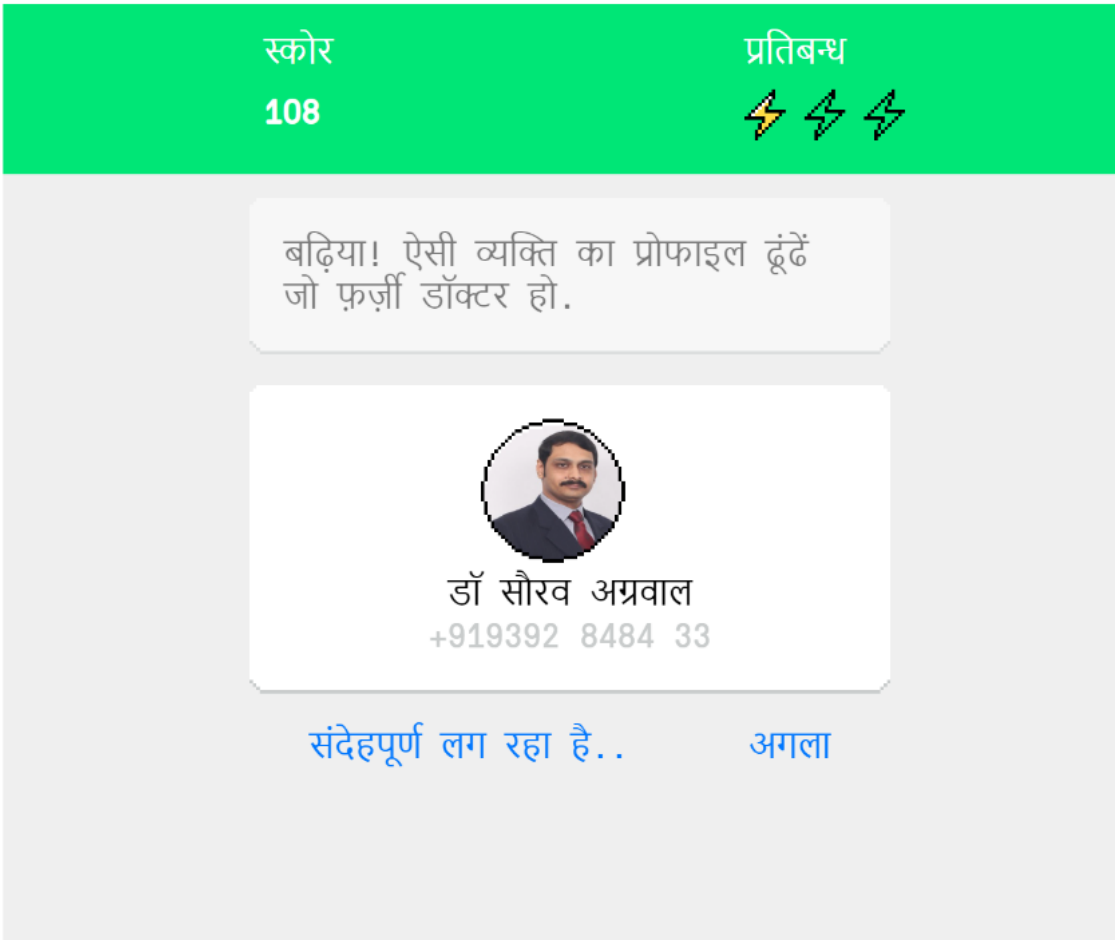


Figure S8 In-game screenshot showing how the Fake News techniques is taught. Continuation of Figure S7. Translation:
Green Bar (Left to Right): "Score" "Sanctions"
Grey Box: "Well done! Find the profile of a person who is a fake doctor"
White Box: "Dr Saurav Agrawal"
Blue Text: "It looks suspicious..." "Next"

Figure S9 Example of a translated manipulative WhatsApp prompt (with English version from another study) intended to show the use of a fake expert.
Screenshot reads: "Hello!
Nowadays it's been very dry.
Even in the rainy season, it does not rain", "Not sure what's happening with the weather these days.
Maybe this is happening because of the climate change in the environment",
"Do you think this is happening because of climate change?",
"I'm not sure, it's difficult to say, farming has become very difficult",
"Hello, I am a scientist, climate change is a big reason for whatever is happening in our environment.
We have to save our environment.",
"Right, interesting".