Countering Misinformation in India through Prebunking

Sushmeena Parihar*[1], Trisha Harjani*[1], Priyank Mathur[2], Beth Goldberg[3], Sander van der Linden[1], & Jon Roozenbeek[4,1,x]

[1] Department of Psychology, University of Cambridge, Cambridge, United Kingdom

[2] Mythos Labs, Delhi, National Capital Region, India

[3] Jigsaw (Google), Google LLC, New York, NY, United States of America

[4] Department of War Studies, King's College London, London, United Kingdom

*denotes equal contribution.

Author note

The studies presented here were funded by Google Jigsaw, Omidyar Network India, and the Alfred Landecker Foundation. All information required to replicate our findings, including item sets, preregistrations, data files, and cleaning and analysis scripts, can be found on our OSF page: https://osf.io/d3zu4/?view_only=9658a09083fe47d1ba5edd84825bf12e.

[x] Corresponding author: jon.roozenbeek@kcl.ac.uk. Department of War Studies, King's College London, London, United Kingdom.

Abstract

Misinformation poses a substantial societal challenge, including in the Global South. However, to date, the vast majority of research into understanding and countering misinformation has taken place in the Global North, and the practical, methodological, and financial complications of doing research with underrepresented groups in Global South countries pose substantial limitations. For this study, we developed five humorous, live-action "prebunking" videos starring a well-known influencer from India, each tackling a different manipulation technique often used on Indian social media: spoofing, decontextualization, the perfect solution fallacy, emotional manipulation, and scapegoating. We hypothesized that the videos would significantly improve people's ability to correctly identify these techniques in news and social media content in an online pilot study (N_1 = 547), two field studies (N_2 = 4,272, N_3 = 1,676), and a large-scale study on YouTube (N_4 = 129,710). We find limited support for our main hypotheses, observing some significant but no consistent improvement in technique discernment. In a series of focus group discussions (N_5 = 33) and an additional randomized controlled trial (N_6 = 827), we explore why this is the case. We find that, while the videos were considered highly entertaining and useful by the majority of viewers, the item rating tasks we administered (evaluating a series of social media posts) were an unfamiliar method of efficacy assessment for our participants. We discuss how interventions can be best designed and measured in contexts where assumptions from the Global North may not apply.

Significance statement

Misinformation in the Global South continues to pose a substantial challenge, but efficacious interventions aimed at reducing individual susceptibility to misinformation remain few and far between, and what "works" in Western countries does not necessarily work elsewhere. For this study, we created and tested five humorous "prebunking" (pre-emptive debunking) videos aimed at building awareness of several manipulation techniques commonly used in misinformation in India: spoofing, decontextualization, the perfect solution fallacy, emotional manipulation, and scapegoating. Across six studies with more than 130,000 participants, we test how well the interventions work at improving people's ability to recognize these techniques, finding limited support for our hypotheses. Our findings have major implications for the design and testing of interventions in the Global South.

Countering Misinformation in India through Prebunking

The spread of misinformation poses a substantial global challenge that has proven to be difficult to tackle at scale (Kozyreva et al., 2024; Ecker et al., 2024; Harjani et al., 2023). Despite many advances in our understanding of the drivers of misinformation belief and spread (Ecker et al., 2022; van der Linden, 2022), as well as how to tackle the problem at the individual and systemic level, a large majority of misinformation research has taken place in the Global North, with research from the Global South facing substantial methodological, practical, financial, and political challenges (Badrinathan & Chauchard, 2023). This is in part due to the fact that Global South research receives comparatively little funding, which limits opportunities to conduct high-quality, large-sample research.

To counter misinformation, researchers have developed both individual-level and system-level interventions. The former can be divided into *nudging* (aimed at changing sharing behavior), *refutation strategies* (debunking and content labelling), and *boosting* (aimed at fostering competences or creating new ones); see Kozyreva et al., (2020, 2024). Prebunking and psychological "inoculation" are types of boosting interventions that seek to pre-emptively build resistance against misinformation, and particularly the persuasion strategies that underlie its effectiveness (van der Linden, 2022). Despite evidence indicating that prebunking and inoculation are efficacious at building resistance against deception (e.g., Roozenbeek et al., 2022), these methods have primarily been tested in Western countries, and evidence on their efficacy in the Global South is lacking (Blair et al., 2024).

We address these problems in the present study by developing and testing five "prebunking" interventions in three Indian states. The primary goal of this project was to develop five videos, each of approximately 1.5-2 minutes in length, that build psychological resistance against manipulation techniques commonly used in misinformation in India. This approach is grounded in a method called "technique-based inoculation" (Roozenbeek & van der Linden, 2024). By "manipulation technique" we here mean rhetorical devices and strategies that intend to exploit, control, or otherwise influence recipients' behaviour, in this case on social media. To be suitable for inoculation in the present context, such techniques must be 1) known to be epistemologically dubious, and 2) be identifiable by any person familiar with the manipulation technique when it is used in written or audiovisual content one might encounter online (Roozenbeek, van der Linden, et al., 2022).

The first phase of this project involved determining which manipulation techniques the videos should cover. In the first section of this paper, we therefore 1) briefly explain "technique-based inoculation", the theoretical foundation of this project, 2) discuss why inoculation videos are a particularly powerful format for tackling misinformation in India, 3) present a brief literature review of misinformation in India, with a view of identifying common manipulation techniques, and 4) present our final selection of five manipulation techniques that are addressed in each of the inoculation videos. Next, we present the results of six separate studies into the efficacy of the five videos in terms of boosting people's ability to spot various manipulation techniques: an online pilot study with the panel provider Respondi (N_1 = 547), two field studies in two different Indian states (N_2 = 4,272, N_3 = 1,676), a large-scale field study on YouTube (N_4 = 129,710) a series of focus group discussions (N_5 = 33), and an adapted randomized controlled trial based on the findings from Studies 1-5 (N_6 = 827).

Transparency and Openness

We describe our sampling plan, all data exclusions (if any), all manipulations, and all measures separately in each study. All data, cleaning and analysis code, and research materials are available at https://osf.io/d3zu4/?view_only=9658a09083fe47d1ba5edd84825bf12e. Data were analyzed using R, version 4.3.3. Visualizations were made using ggstatsplot, version 0.12.3 (Patil, 2021), and ggplot2, version 3.5.0 (Wickham, 2016). Studies 1, 2, 3, and 4 were preregistered (see the individual studies for preregistration links).

Inoculation theory and "technique-based inoculation"

Inoculation theory is a framework from the 1960s which posits that it is possible to build preemptive psychological resistance against future unwanted persuasion attempts (Compton, 2013; McGuire & Papageorgis, 1961). Much like a medical vaccine, where people are injected with a weakened or dead virus or other pathogen which triggers the immune system to produce antibodies, psychological inoculations aim to prevent unwanted persuasion from happening in the first place ("prebunking"), rather than correct it post-exposure ("debunking"). Over many decades of research, inoculation has proven to be effective at increasing resistance against unwanted persuasion attempts (Banas & Rains, 2010; Traberg et al., 2022; Lu et al., 2023).

Traditionally, inoculation interventions have sought to build resistance against *specific* fallacious arguments, for example in the context of climate change misinformation (van der Linden et al., 2017) and the detrimental health effects of smoking (Pfau et al., 1992). This so-called *issue-based inoculation* approach is useful in the context of misinformation when it is possible to predict with reasonable certainty what kind of misinformation people will be exposed to in the near future. For instance, before the 2020 US presidential elections, Twitter launched a so-called “prebunking” campaign against misinformation about election fraud (Ingram, 2020); because accusations of election fraud are a recurring theme on US social media during controversial elections, it was predictable that such narratives would make the rounds online, which meant that Twitter could design a campaign countering these narratives before they went viral.

In many cases, however, it is not possible to predict what specific misinformation people will likely be exposed to. Nonetheless, this does not mean that it becomes impossible to inoculate people against misinformation altogether. Much of the misinformation that goes viral online makes use of a limited number of manipulation techniques or tropes, which are indicators of low epistemic quality (Carrasco-Farré, 2022; Simchon et al., 2021). Such manipulation techniques include, for example, logical fallacies, conspiratorial reasoning, emotional manipulation, trolling, and false amplification (Cook et al., 2022; van der Linden & Roozenbeek, 2020). Previous research has shown that people can be successfully inoculated against the use of these techniques in social media content and news headlines that they are entirely unfamiliar with (Basol et al., 2021; Harrop et al., 2023). This approach of inoculating people against manipulation techniques rather than individual arguments is commonly referred to as *technique-based inoculation*.

Using videos to inoculate against misinformation in India

A growing body of work has found that video is a particularly powerful medium for technique-based inoculation interventions (Hughes et al., 2021; Piltch-Loeb et al., 2022). The advantage of using video over other types of interventions such as games (Roozenbeek & van der Linden, 2019) is that they can be easily scaled across social media platforms. Roozenbeek, van der Linden et al. (2022), for instance, found that showing inoculation videos as advertisements on YouTube boosted viewers’ ability to correctly identify manipulation techniques in previously unseen news headlines by about 5-10%, compared to a control group. This finding showed that it is feasible to improve people’s ability to identify manipulative content even in a noisy social media environment, where people do not have to pay attention to the video, can turn off the sound, switch to another tab, or disengage from the intervention in some other manner (Jigsaw, 2024).

India is YouTube's largest and fastest growing market worldwide (The Hindu, 2019). After the government banned TikTok in June 2020, India experienced a massive boom in home-grown video content production. Today, YouTube is the most popular source of news in India, with 53% of Indians reporting that they use the platform to access news content (Basuroy, 2022). At the same time, misinformation has become a growing problem on Indian social media (Badrinathan, 2021). As the number of social media users continues to grow, so does the amount of harmful, false, and/or misleading viral content. Countering this problem can be difficult, as educational interventions have in the past failed to yield an improvement in Indian social media users' ability to identify misinformation (Badrinathan, 2021). In addition, content moderation and debunking efforts are often not feasible, for example because direct messaging apps such as WhatsApp and Telegram are encrypted, making it impossible to know exactly what kind of misinformation is shared and by who (Pasquetto et al., 2020). Therefore, to leverage the viral potential of video content, the goal of this project is to produce videos that improve resilience against the types of manipulative content that Indian social media users are likely to encounter. These videos can then be shared on encrypted direct messaging apps, or run as ads on YouTube, so as to show them to as many people as possible.

Conducting intervention studies in India

Viral misinformation in India is has had serious potential adverse consequences (Vasudeva & Barkdull, 2020; Arun, 2019). Past research has yielded video-based interventions that "inoculate" people against manipulation techniques known to be common in Global North contexts (although knowing exactly how common each technique is vis-a-vis other techniques is a highly complex task, see Coan et al., 2021). Roozenbeek, van der Linden et al. (2022) developed a series of videos that tackled the following manipulation techniques: 1) emotionally manipulative language intended to evoke negative emotions such as fear, anger or outrage; 2) the use of incoherent or mutually exclusive arguments; 3) false dichotomies (or false dilemmas); 4) scapegoating or holding an individual or group solely responsible for a complex problem with multiple causes; and 5) engaging in ad-hominem attacks (i.e., attacking the person instead of engaging with the argument). However, it is not necessarily the case that these manipulation techniques are commonly encountered outside of

Western contexts. It is therefore of key importance to survey what tactics and manipulation techniques are common in misinformation found on Indian social media.

Mythos Labs, an India-based company which leverages comedy to produce educational interventions such as videos and educational programs (https://mythoslabs.org/) investigated the proliferation of misinformation in India in January 2023. Their report identifies three common tactics used to spread misinformation: *spoofing*, *decontextualisation,* and *social proofing*. *Spoofing* refers to false or misleading content that mimics the look and feel of reliable sources. Such content can come from parody accounts (which imitate well-known individuals) or misattribute claims to influential figures. This tactic is sometimes also referred to as “impersonation” (van der Linden & Roozenbeek, 2020).

Decontextualisation consists of removing or editing content to distort the facts, for example by cropping an image so that important information is left out, misrepresenting the time, date or location of real information, or leaving out crucial contextual information about a particular news story.

In India, information sometimes proliferates on social media about fake cures and remedies; this type of content can be said to fall under the “fake expert” manipulation technique (Basol et al., 2021; Cook et al., 2022). In addition, health misinformation tends to amplify fears and concerns that people may have about a particular disease or treatments such as vaccines (Al-Zaman, 2022). The underlying fallacy here is the *perfect solution fallacy*, or the idea that simple solutions exist to complex issues that (in this case medical) science has been unable to solve.

Misinformation often seeks to amplify negative sentiments about groups or individuals, disregarding their actual responsibility or culpability. In the literature, this type of content (where people or groups are singled out as being responsible for a complex problem) is referred to as *scapegoating* (Atlani-Duault et al., 2020).

Finally, one common technique used in a broad range of misinformation is *emotional manipulation* (Roozenbeek, van der Linden, et al., 2022). Evoking negative emotions in particular are key to the spread of (mis)information on social networks (Brady et al., 2017; Vosoughi et al., 2018). Content that seeks to evoke strong emotions such as fear, anger or outrage is exceedingly common, and can override considerations of accuracy (Van Bavel et al., 2021); this is known as the “appeal to emotion” fallacy (CITE).

Selecting manipulation techniques for inoculation

These sources suggest that these five manipulation techniques are commonly encountered on Indian social media and would therefore be suitable to develop inoculation videos to educate Indian users about these techniques:

- **Spoofing (or impersonation)**: imitating or mimicking real individuals, in order to drive engagement through borrowed credibility.
- **Decontextualization**: leaving out crucial information in order to make an event or story look more outrageous than it really is.
- **Perfect Solution Fallacy**: promoting false cures and remedies to various diseases in order to create false hope or fear, to persuade people to buy something.
- **Emotional manipulation**: seeking to evoke strong emotions such as fear, anger or outrage, with a view of overriding people's considerations of accuracy.
- **Scapegoating**: holding individuals or groups responsible for adverse events (real or imaginary) to instigate conflict.

These five manipulation techniques do not comprise *all* misinformation going viral in India. The videos were created in collaboration with Indian content creators and well-known actors, who starred in the videos. They were originally shot in Hindi and subsequently dubbed into Marathi and Malayalam. See Appendix I for links to the videos (on YouTube).

Study 1: Online Pilot

As a first test, we conducted a pilot study with the "spoofing" video using the online recruitment panel Respondi. In total, we recruited 547 participants from India. The goal of this study was to assess the initial efficacy of the "spoofing" video as a way to improve people's ability to recognise manipulative messaging in social media content (in this case direct messaging apps). Taken together, the results are partly successful: the "spoofing" video significantly and meaningfully boosted recognition of manipulative messaging, but also had a (marginally significant and substantially smaller) effect on people's evaluation of *non*-manipulative messaging, that is, "real news". Overall, the effect on the video on "discernment", that is, the ability to distinguish manipulative from non-manipulative messaging, is not significant but trending in the right direction..

Methods

We conducted a preregistered randomised controlled trial on Respondi (an ISO-certified online panel provider) with 547 participants from India (preregistration link: https://osf.io/nsp7a/?view_only=488d3cec255d402b97ce6153b1a084a4, anonymized for peer review). Participants were randomly assigned to a treatment (spoofing video) or control (unrelated video) condition. Next, participants were shown a series of 10 WhatsApp posts (stripped of identifying information), 5 of which made use of spoofing, and 5 of which contained true information phrased in a neutral manner ("real news"). Participants were asked to evaluate each of these posts using the following outcome measures (1 being "strongly disagree" and 5 being "strongly agree"): 1) this message contains spoofing (**technique recognition**); 2) I am confident that this message contains spoofing (**confidence**); 3) I would forward this message to others on direct messaging apps (**sharing**); and 4) I trust this post (**trust**). Performance was calculated by taking the average of the 5 manipulative (spoofing) items and the 5 real news (real) items, and then calculating the difference score ("discernment") for the technique recognition, sharing, and trust measures. The survey was conducted in English throughout. We hypothesised the following:

H1: Participants in the treatment (inoculation) group are significantly better than the control group at discriminating WhatsApp content that contains spoofing from content that does not.

H2: Participants in the treatment group are significantly more confident in their ability to discriminate the manipulativeness of manipulative and neutral social media content.

H3: Participants in the treatment group have significantly better sharing discernment (a measure of the quality of their online content sharing decisions), compared to a control group.

H4: Participants in the treatment group are significantly better than the control group at discriminating the trustworthiness of WhatsApp content that contains spoofing and content that does not.

Results

Main Analyses

For **H1**, we find that the treatment group is significantly better than the control group at correctly identifying content containing spoofing ($p < .001$, $d = .32$). In addition, we find a marginally significant effect on “real news”, so that treatment group participants are more likely to identify “real news” as containing spoofing (*p* = .04, *d* = .18). Overall, this leads to non-significant discernment (*p* = .18, *d* = .12), although the results trend in the right direction. We thus find no support for **H1**. See Figure 1.

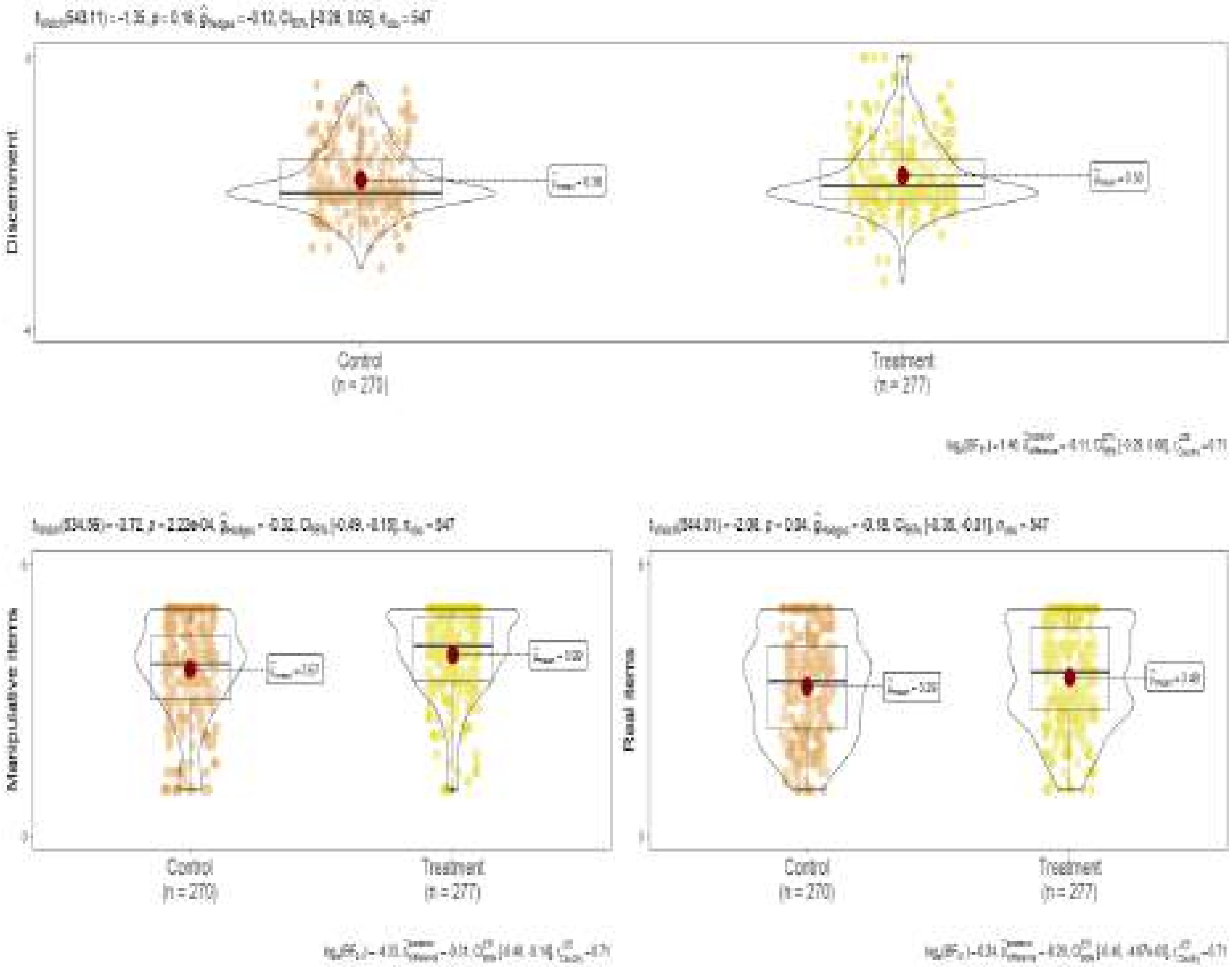

Figure 1. Box-violin plots with data jitter for treatment and control group participants for the “technique recognition” measure, for discernment (top panel), manipulative items (bottom left), and “real news” (bottom right). Statistical tests (including Bayesian t-tests) are provided in the panels.

Second, for **H2**, we find that treatment group participants are significantly more confident than the control group in their assessment of whether WhatsApp messages contain spoofing (*p* < .001, *d* = .34). Similarly, participants do not become more confident that “real news” contains spoofing (which was as expected, *p* = .058, *d* = .16). These results support **H2**. For **H3** and **H4**, we find no significant differences between the treatment and control groups, neither for the forwarding nor trust measures (and neither for the manipulative nor the real items or discernment), all *p* values > .177. These results fail to confirm **H3** and **H4**. See Table 1.

Variable	**Statistic**	**df**	***p***	**M_{diff}**	**Cohen's *d***
Recognition	-3.723	545	**< .001**	-0.322	-0.318
Confidence	-3.932	545	**< .001**	-0.321	-0.336
Forwarding	1.083	545	0.279	0.123	0.093
Trustworthiness	0.867	545	0.386	0.101	0.074
Recognition (real news)	-2.085	545	**0.038**	-0.203	-0.178
Confidence (real news)	-1.903	545	0.058	-0.181	-0.163
Forwarding (real news)	0.807	545	0.420	0.081	0.069
Trustworthiness (real news)	1.085	545	0.279	0.108	0.093
Recognition Discernment	-1.353	545	0.177	-0.120	-0.116
Forwarding Discernment	-0.562	545	0.574	-0.042	-0.048
Trustworthiness Discernment	0.090	545	0.929	0.007	0.008

Table 1. Independent-samples *t*-tests for the technique recognition, confidence, forwarding, and trust measures.

Item-level comparisons

It is possible that item-specific effects are at play. We therefore take a closer look at item effects in Table 2 below, which shows independent-samples *t*-tests at the item level (for the manipulative and real items). The table shows that treatment group participants were significantly more likely than the control group to say that all 5 manipulative items contained spoofing (as hypothesised), all *p* values < .019. For the real items, treatment group participants were significantly more likely to say that the "heatwaves" item ("India to get heat waves this year after hottest February on record!") contained spoofing (*p* = .002). A smaller effect was found for the "turtles" item ("India and Australia compete for the World Test Championship from Thursday", *p* = .039). It is possible that these items were worded in a non-neutral way, so that treatment group (incorrectly) inferred that they used spoofing.

Variable	**Statistic**	**df**	***p***	**M_{Diff}**	**Cohen's *d***
Manipulative items					
Cricket	-2.403	545	**0.017**	-0.258	-0.205
CEO Phone Company	-2.805	545	**0.005**	-0.314	-0.240
Textile Toxins	-3.715	545	< **.001**	-0.402	-0.318
Bankers Crisis	-3.395	545	**0.001**	-0.370	-0.290
Alcohol Epidemic	-2.354	545	**0.019**	-0.267	-0.201
Real items					
Heatwaves	-3.047	545	**0.002**	-0.370	-0.261
Cricket	-1.044	545	0.297	-0.136	-0.089
Yoga	-0.477	545	0.634	-0.060	-0.041
Millets	-1.732	545	0.084	-0.211	-0.148
Turtles	-2.067	545	0.039	-0.236	-0.177

Table 2. Independent samples *t*-tests at the item level on technique recognition (H1). Red indicates a significant effect in the opposite direction than hypothesized.

Discussion

Overall, we find that the "spoofing" video does what it was designed to do, namely, increase people's ability to identify the use of spoofing in messages that contain spoofing. In addition, people become more confident in their ability to identify spoofing. However, we also find smaller effects of the video on messaging that does *not* make use of spoofing, possibly because people become more sceptical overall, or because the items were designed incorrectly and were too ambiguous. Due to the weaker effects of the video on "real news" than the manipulative items and the fact that only one item ("heatwaves") appears to have been primarily responsible for this effect, we primarily recommend reconsidering design decisions of the item sets.

Furthermore, we find no effect of the video on the "forwarding" and "trust" measures, with *p*-values for discernment being .573 and .929, respectively; these measures do not approach significance, and it is possible that this indicates the true absence of an effect (rather than non-significance due to low statistical power or insufficient sample size). We return to this discussion in Study 5.

Study 2: Field Study (Hindi)

Through our survey providers, Outline India, we conducted an in-person field study with 4,272 participants (roughly *n* = 800 across the treatment and control conditions for each of the five videos (these are linked and described in the *Introduction* section) in a predominantly Hindi-speaking Indian state (Uttar Pradesh). Specifically, we recruited 671 participants for the *spoofing* video, 895 for the *decontextualization* video, 910 for the *perfect solution fallacy*, 895 for *emotional manipulation*, and 894 for *scapegoating*; the total number of participants was thus *N* = 4,272. Studies 2 and 3 (see below) were preregistered here: https://osf.io/257u6/?view_only=7899f050fa43414ba1fc6c5e2f0943e2 (link anonymized for peer review). We preregistered a target sample size of *N* = 4,400, leaving us somewhat short of our target. We also note that due to an administrative error on our part, we failed to publish this preregistration prior to data collection (it was left as a draft on the OSF). However, we did not change any of our procedures after data collection.

Methods

After indicating consent, participants were randomly assigned to a treatment or control condition. They were then asked to watch a video on a tablet provided by Outline India. Participants in the treatment condition saw one of the five inoculation videos and those in control condition saw an unrelated video of similar length. Next, and similar to the pilot study, participants were shown a series of 10 WhatsApp posts (stripped of identifying information), 5 of which made use of the manipulation technique from the prebunking video that was shown, and 5 of which contained true information phrased in a neutral manner ("real news"); the wording of these items was changed slightly compared to the pilot study. See our OSF page for item details: . Participants were asked to evaluate each of these posts using the following outcome measures: 1) this post contains spoofing/decontextualization/perfect solution fallacy/emotional manipulation/scapegoating (**technique recognition**); 2) How confident are you about your previous answer (**confidence**); and 3) How likely are you to forward this post to others on WhatsApp? (**sharing**); Performance was calculated by taking the average of the 5 manipulative (technique utilizing) items and the 5 real news (real) items, and then calculating the difference score ("discernment") for the technique recognition and sharing measures. We hypothesised the following:

H1: Participants in the treatment (inoculation) group are significantly better than the control group at discriminating WhatsApp content that contains spoofing/decontextualization/perfect solution fallacy/emotional manipulation/scapegoating from content that does not.

H2: Participants in the treatment group are significantly more confident in their ability to discriminate the manipulativeness of manipulative and neutral social media content.

H3: Participants in the treatment group have significantly better sharing discernment (a measure of the quality of their online content-sharing decisions), compared to a control group.

Results

Spoofing

For **H1**, we find that the treatment group is significantly better than the control group at correctly identifying content containing spoofing (*p* = .003, *d* = .23). In addition, we find a significant positive effect on "real news", so that treatment group participants are more likely to identify "real news" as *not* containing spoofing (*p* < .001, *d* = .32). Overall, this leads to significant discernment (*p* < .001, *d* = .38), as treatment group participants are more accurate than the control group at correctly identifying both spoofing-content and "real news". This provides support for **H1**. See Figure 2.

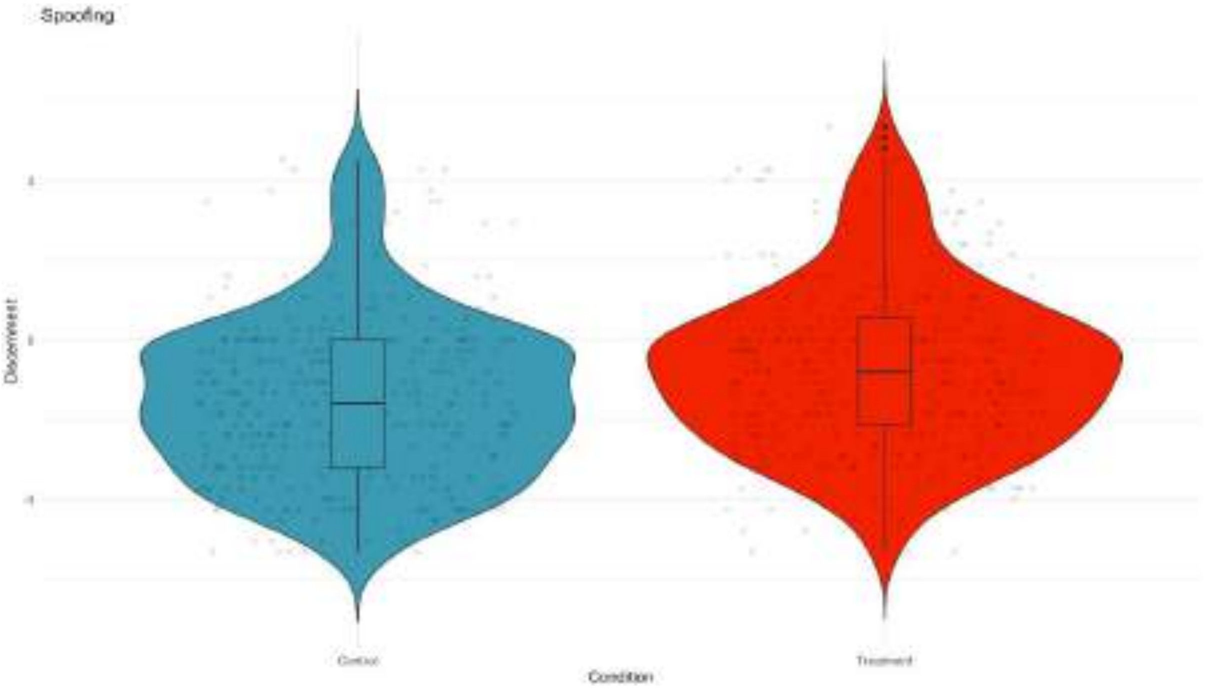

Figure 2. Box- violin plot showing participants' ability to discern content containing spoofing from "real news" (discernment).

For **H2** and **H3**, we find no significant differences between the treatment and control groups, neither in their assessment of whether WhatsApp messages contain spoofing nor the forwarding measure. These results fail to confirm **H2** and **H3**. See Table S1 in the Supplementary Information.

Decontextualization

For **H1**, we find that the control group is significantly better than the treatment group at correctly identifying content containing decontextualization (p = .012, d = .17); this result is in the opposite direction than hypothesized. In addition, and as hypothesized, we find a significant effect on "real news", so that treatment group participants are *less* likely to identify "real news" as containing decontextualization (p = .001, d = .23). Overall, this leads to non-significant discernment (p = .58, d = .04). These results fail to confirm **H1**. Second, for **H2**, we find that treatment group participants are significantly more confident than the control group in their assessment of whether WhatsApp messages contain decontextualization (p < .001, d = .25). However, treatment group participants do not become more confident that "real news" contains decontextualization (p = .085, d = .12). These results support **H2**. For **H3**, we find that the control group participants are significantly more likely than the treatment group to forward real news to others (p = .003, d = .20). Overall, this leads to non-significant discernment (p = .08, d = .12). These results fail to confirm **H3**. See Table S2 in the Supplementary Information.

Perfect Solution Fallacy

For **H1**, we find no significant difference between the treatment and control groups at correctly identifying content containing a perfect solution fallacy (p = .67, d = -.028). In addition, we find a significant effect on "real news", such that the treatment group participants are less likely to indicate that "real news" contains a perfect solution fallacy (p = .002, d = .21). This leads to a significant discernment (p = .031, d = .14), providing support for **H1**. For **H2**, we find that treatment group participants are significantly more confident than the control group in their assessment of whether WhatsApp messages contain perfect solution fallacy (p < .001, d = .45) as well as "real news" (p = .002, d = .20). These results support **H2**. For **H3**, contrary to our hypothesis, we find that the treatment group participants are significantly more likely than the control group to forward WhatsApp messages containing a perfect solution fallacy (p = .010, d = .17), but no significant difference in sharing intentions of "real news" (p = .966, d = -.003). Overall, this leads to significant differences in sharing discernment in the *opposite* direction than hypothesized (p = .002, d = .21). These results fail to support **H3**. See Table S3 in the Supplementary Information.

Emotional Manipulation

For **H1**, we find no significant differences between the treatment and control groups in their ability to correctly identify content containing emotional manipulation versus “real news”; all p values > .40. These results fail to confirm **H1**. Second, for **H2**, we find that treatment group participants are significantly more confident than the control group in their assessment of whether WhatsApp messages contain emotional manipulation (p < .001, d = .51) as well as “real news” (p = .001, d = .22). These results support **H2**. For **H3**, we find no significant differences between the treatment and control groups for the forwarding measure (and neither for the manipulative nor “real news” or discernment), all p values > .08. These results fail to confirm **H3**. See Table S4 in the Supplementary Information.

Scapegoating

For **H1**, we find that the control group is significantly better than the treatment group at correctly identifying content containing scapegoating (p < .001, d = .38). Overall, we find a significant discernment in the opposite direction than hypothesized (p < .001, d = .30). These results fail to confirm **H1**. Second, for **H2**, we find that treatment group participants are marginally significantly more confident than the control group in their assessment of whether WhatsApp messages contain scapegoating (p = .049, d = .13). However, participants do not become more confident that “real news” contains scapegoating (p = .23, d = .08). These results partially support **H2**. For **H3**, we find no significant differences between the treatment and control groups for the manipulative messages (p = .522, d = -.043) and “real news” (p = .097, d = .111). However, we find that control group participants have significantly higher sharing discernment, contrary to our hypothesis (p = .006, d = .188). These results fail to confirm **H3**. See Table S5 in the Supplementary Information.

Discussion

In this large-scale field evaluation (total N = 4,272), we have tested the effectiveness of five “prebunking” videos along three outcome measures: manipulation technique recognition, confidence (attitudinal certainty), and intentions to share misleading and neutral (“real”) content. We summarize the results in Table 3.

	Outcome measure		
	Technique recognition	**Confidence**	**Sharing**
Spoofing	Significant hypothesized effect for spoofing items & discernment	No significant differences	Marginal hypothesized effect for "real news"
Decontextualization	Significant *opposite* effect for decontext. items, sig. effect for real items	Significant hypothesized effect for decontextualization items, not real news	Significant hypothesized effect for real news
Perfect Solution Fallacy	Significant hypothesized effect for real items & discernment	Significant hypothesized effect for both fallacy & real items	Significant *opposite* effect for fallacy items & discernment
Emotional Manipulation	No significant differences	Significant hypothesized effect for both manipulative & real items	No significant differences
Scapegoating	Significant *opposite* effect for scapegoating items & discernment	Significant hypothesized effect for scapegoating items	Significant *opposite* effect for discernment

Table 3. Summary of the results for each video. Green color indicates that the main hypothesis was supported; orange indicates mixed support; red indicates a backfire effect, i.e., the results are significant but in the opposite results than hypothesized.

Our findings indicate that the “spoofing” video works best in fulfilling its intended objective of increasing people’s ability to identify the use of spoofing in messages that contain the technique. The “perfect solution fallacy” video ranks as the second most effective in terms of technique recognition. We also find that people become more confident in their ability to identify “perfect solution fallacy” and emotional manipulation.

However, contrary to previous research (Capewell et al., 2024; Leder et al., 2024; Maertens et al., 2024; Roozenbeek, van der Linden, et al., 2022), we also find evidence of backfire effects: for the decontextualization and scapegoating videos, treatment group participants were significantly worse than control group participants at correctly distinguishing between misleading and neutral content. This is a rare finding in the inoculation/prebunking literature (Roozenbeek et al., 2023; 2024), which tends to find either significant effects in the hypothesized direction or no between-group differences.

Why this is the case is difficult to untangle using the data collected for this study. One possibility is methodological limitations: survey participants may not have been used to the type of survey that was administered (our implementation partners indicated this may have been the case), which may have limited the validity of their responses. In addition, participants may have been confused or bored with the survey, leading to repetitive response patterns. We address these possibilities in the next study.

Study 3: Field Study (Malayalam)

Study 2 was conducted in Uttar Pradesh, a Hindi-speaking state, which provides limited generalizability. Through our survey providers, Outline India, we therefore conducted an in-person field study with 1,676 Malayalam-speaking participants from Kerala, a state in the south of India. Due to budget limitations, we were unable to include all five videos tested in Study 2. Instead, we chose to focus on two videos that showed the most optimistic findings (decontextualization and scapegoating). Across the treatment and control conditions, we recruited 838 participants for the former and 838 for the latter. We preregistered this study here, together with Study 2: https://osf.io/257u6/?view_only=7899f050fa43414ba1fc6c5e2f0943e2 (link anonymized for peer review). We preregistered a sample size of *N* = 1,600, which means we recruited slightly more participants than originally intended.

Methods

Following Study 2, after indicating their consent, participants were randomly assigned to treatment or control condition. They were then asked to watch a video on a tablet provided by Outline India. Participants in the treatment condition either saw the decontextualization or scapegoating inoculation video and those in control condition saw an unrelated video of similar length. Next, participants were shown the same series of 10 WhatsApp posts from Study 2, translated into Malayalam (stripped of identifying information), 5 of which made use of the manipulation technique from the prebunking video that was shown, and 5 of which contained true information phrased in a neutral manner (“real news”). Participants were asked to evaluate each of these posts using the following outcome measures: 1) this post contains decontextualization/scapegoating (**technique recognition**); 2) How confident are you about your previous answer (**confidence**); and 3) How likely are you to forward this post to others on [direct messaging apps]? (**sharing**); Performance was calculated by taking the average of the 5 manipulative (technique utilizing) items and the 5 real news (real) items, and then calculating the difference score (“discernment”) for the technique recognition and sharing measures. Our hypotheses were the same as for Study 2.

Results

Decontextualization

For **H1**, we find that the control group and the treatment group are not significantly different at correctly identifying content containing decontextualization ($p = .14$, $d = .10$). In addition, and as hypothesized, we find a significant effect on "real news", so that treatment group participants are *less* likely to identify "real news" as containing decontextualization ($p = .002$, $d = .21$). Overall, this leads to significant discernment ($p = .001$, $d = .22$); this result is in the opposite direction than hypothesized. These results fail to confirm **H1**. Second, for **H2**, we find that the control group participants are significantly more confident than the treatment group in their assessment of whether messaging app messages contain decontextualization ($p < .001$, $d = .39$) as well as for "real news" ($p < .001$, $d = .28$). These results fail to confirm **H2**. For **H3**, we find no significant differences between the treatment and control groups for the forwarding measure (and neither for the manipulative nor "real news" or discernment), all *p* values > .57. These results fail to confirm **H3**. See Table S6 in the Supplementary Information.

Scapegoating

For **H1**, we find that the control group is significantly better than the treatment group at correctly identifying content containing scapegoating ($p < .001$, $d = .29$). Although, as postulated, we find a significant effect on "real news", such that the treatment group participants are more likely to correctly identify "real news" ($p = .005$, $d = .19$). However, this leads to a significant discernment in the opposite direction ($p < .001$, $d = .34$). Overall, these results partial support for **H1**. Second, for **H2**, we find that control group participants are significantly more confident than the treatment group in their assessment of whether messaging app messages contain scapegoating ($p < .001$, $d = .46$) as well as "real news" ($p < .001$, $d = .35$). These results fail to confirm **H2**. For **H3**, contrary to our hypothesis, we find that the treatment group participants are significantly more likely than the control group to forward messaging app messages containing scapegoating ($p = .004$, $d = .20$) but are significantly more likely than the control group to forward real news to others ($p = .001$, $d = .24$). Overall, this leads to non-significant discernment ($p = .33$, $d = .07$). These results fail to confirm **H3**. See Table S7 in the Supplementary Information.

Discussion

In this replication study (total $N = 1,676$), we tested the effectiveness of the decontextualization and scapegoating "prebunking" videos along three outcome measures:

manipulation technique recognition, confidence (attitudinal certainty), and intentions to share misleading and neutral ("real") content. We summarize the results in Table 4.

	Outcome measure		
	Technique recognition	**Confidence**	**Sharing**
Decontextualization	Significant *opposite* effect for discernment, sig. hypothesized effect for real items	Significant *opposite* effect for decontext. and "real news" items	No significant effects
Scapegoating	Significant *opposite* effect for scapegoating items and discernment, sig. hypothesized effect for real items	Significant *opposite* effect for scapegoating and "real news" items	Significant *opposite* effect for scapegoating items, sig. hypothesized effect for real items

Table 4. Summary of the results for each video. Red indicates a backfire effect, i.e., the results are significant but in the opposite results than hypothesized.

Unlike Study 2, which shows some degree of success for the "spoofing" and "perfect solution fallacy" videos in the previous study, the results from this study indicate that neither the "decontextualization" nor the "scapegoating" video is successful in fulfilling its intended objectives of (1) increasing people's ability to identify the use of decontextualization/scapegoating in messages that contain the technique, (2) increase people's confidence levels in their ability to discriminate the manipulativeness of manipulative and neutral social media content, and (3) increasing people's sharing discernment (a measure of the quality of their online content-sharing decisions). We discuss our findings in more detail in the Discussion and Conclusion section.

Study 4: YouTube field study

Following Roozenbeek, van der Linden et al. (2022), we ran a YouTube ad campaign with the five videos across three languages (Hindi, Marathi, and Malayalam). The videos were shown as YouTube ads to a total of 68,305,284 Indian YouTube users across the latter half of 2023, of whom 24,054,082 watched the video for at least 30 seconds (which YouTube counts as a "video view"). Then, using the YouTube "BrandLift" survey tool, some users who had seen the video were shown one of several single-item multiple-choice survey questions which assessed their ability to recognize the manipulation technique from the video they watched in a (fictitious) headline. This occurred within 24 hours of watching the ad, at the end of a different YouTube video they chose to watch (participants were free to ignore the question). For example, some participants were asked to evaluate the following sentence: "*ALL NATURAL REMEDY: Chikoo skin extracts removes all dengue symptoms within hours!*", and given four possible response options: 1) a command, 2) scapegoating, 3) fake cure (correct option), and 4) none of the above. We developed three such survey items per video and translated them to Hindi, Marathi, and Malayalam. All videos (and hence all survey items) were deployed in Hindi; for Marathi and Malayalam, we only implemented the Spoofing, Fake Cures, and Emotional Manipulation videos (and hence survey items). In tandem with this, a control group of Indian YouTube users (similar demographically to the treatment group) was also shown a single survey question without having been shown any inoculation videos as an ad. This allowed us to compare the percentage of correct responses (25% being chance level) between YouTube users who had seen an inoculation video versus those who had not. In total, we collected n = 65,051 survey responses for Hindi, n = 35,066 for Marathi, and n = 29,593 for Malayalam (see Table S8 for details). This means the total sample size was N = 129,710. We preregistered this study (see our OSF page for the anonymized preregistration from AsPredicted.org). Our procedure and analyses follow exactly those used by Roozenbeek, van der Linden, et al. (2022).

Results

The results of the two-proportion z-tests (comparing the proportions of correct responses in the treatment and control group; Roozenbeek, van der Linden, et al., 2022). We found significant effects in the hypothesized direction for a minority of items in all languages (2/15 items for Hindi, 2/9 items for Marathi, and 5/9 items for Malayalam). Combining all items together for each language, we find that Hindi-speaking participants were slightly better than the control group (p = .003), as were Malayalam speakers (p < .001), but this was not the case for Marathi (p = .157). Effect sizes were low: Cohen's h = .022 for Hindi and h = .049 for Malayalam, descriptively lower than the h = ~.09 improvement reported by Roozenbeek, van der Linden, et al. (2022), who also found significant between-group differences in 4 out of 6 items (as well as a significant overall effect). See Table S9 for a full overview.

Discussion

In this study, we conceptually replicated the YouTube field study by Roozenbeek, van der Linden, et al. (2022), using a large sample of nearly 130,000 Indian YouTube users. Overall, we find mixed support for our hypothesis that watching an inoculation video as a

YouTube ad significantly improves one's ability to correctly identify manipulation techniques in news headlines: while the overall effect is significant in two out of three languages (Hindi and Malayalam), we only find significant between-group differences fora minority of items administered, despite a very large sample size. In addition, the overall effect size is small (Cohen's h = ~.05, i.e., well below the minimum threshold for small effect sizes in psychology and behavioral science of Cohen's d (or h) = .20). These findings are in contrast with the study we conceptually replicated (Roozenbeek, van der Linden et al., 2022), who did observe more robust effects ($p < .001$, $h = .09$, with a preregistered smallest effect size of interest of $h = 0.10$) using a highly similar design with inoculation videos administered to YouTube users in the United States. Overall, we therefore conclude that, similar to Studies 1-3, we did not find robust evidence that watching the inoculation videos prompted a significant boost in viewers' ability to correctly identify manipulative or misleading content. However, contrary to Studies 1-3, in this large real-world social media test we do not find evidence that the videos lead to *worse* identification either. As mentioned above, this is possibly due to methodological issues (for instance because the item rating task was considered confusing or difficult), or because the video did not convey information in a way that participants understood what was subsequently expected of them. We explore these possibilities in Studies 5 and 6.

Study 5: Focus Group Discussions

For Study 5, we conducted a series of focus group discussions (FGDs), in order to better comprehend Indian internet users' perceptions of the five inoculation videos (e.g., whether they enjoyed watching them, believed they addressed an important problem, could be readily understood, were effective at getting across their intended message, and so on), as well as the impact these videos have on local users' understanding of, and vulnerability to, online manipulation techniques. The FGDs were carried out by Ormax Media (https://www.ormaxmedia.com/), with help from Mythos Labs.

Methodology

A total of 33 participants took part in six FGDs conducted virtually on Zoom (2 FGDs per age range, one for men and one for women). Participants represented a mix of male and female internet users aged 18 - 44 across metropolitan cities (population over 7.5 Million), urban areas (population between 1 Million - 7.5 Million) and semi-urban areas (population less than 1 Million). All participants were located in one of the following Hindi Speaking

Markets (HSM): Mumbai, Delhi, Kolkata, Madhya Pradesh, Uttar Pradesh, Gujarat, Maharashtra, Rajasthan, Punjab, Haryana, Chandigarh and Himachal Pradesh. See Table 5 for an overview of the sample.

Variable	Category	Count
Age	18-24	10
	25-34	12
	35-44	11
Gender	Female	17
	Male	16
Location	Metropolitan (population > 7.5m)	10
	Urban (population 1m - 7.5m)	15
	Semi-urban (population < 1m)	8

Table 5. Study 5: Sample overview.

Each FGD lasted approximately one hour. All FGDs were moderated by experienced moderators from Ormax Media who had no involvement in the making or dissemination of the videos. Specifically, the FGDs were aimed at understanding what people thought about the videos, how they interpreted them (for example in terms of usefulness and entertainment value), as well as the item rating tasks administered in Studies 1-4 (see above). See Appendix II for the complete Moderator Guide.

Results

We discuss the results from the focus groups by addressing, separately, 1) participants' understanding of the need for the videos, 2) the quality of the videos, and 3) impact of the videos on participants' ability to identify manipulation. In the interest of brevity, we discuss general findings here; for video-specific findings, see Appendix III.

First, all participants said they assume the goal of the videos is to raise awareness of misinformation techniques and to make internet users less vulnerable to misinformation. All participants believed these videos were "needed" and that they would share them with their friends and family. However, some participants expressed confusion as to the difference between an manipulation technique used to spread misinformation, and misinformation itself. This was especially the case among participants above the age of 25. Some participants also said they were not sure if Indian internet users, especially in rural areas, were ready to learn about specific misinformation techniques because they might not even be aware that misinformation exists and is rampant in the first place. They suggested creating more introductory videos explaining the basic concept of misinformation, before showing audiences more advanced videos about identifying specific techniques.

Second, with respect to the quality of the videos, all participants said they like the use of comedy in the videos and that they struck the right balance of humor and information. Some comments included "the tone was good; not too silly and not too dry", "the use of comedy was appreciated", and "humor was good, not too much and not too little". Participants appreciated the rural setting of the videos, saying this would likely make them appealing to the broadest segment of internet users in India. Further, most participants under the age of 34 recognized Saloni Gaur (the star of the videos), but most participants over 34 did not know who she was. Finally, most participants found the dynamic of a young girl explaining concepts to her older uncle to be "interesting", "unique" or "refreshing". However, two female participants said they found it "annoying" that the young girl was explaining things to her older uncle.

Third, "emotional manipulation", "spoofing", and "fake cures" were the easiest manipulation techniques to understand. 100% of participants who saw these videos were able to correctly identify these techniques using a different methodology from the one used in Studies 1-4 (guided by a moderator who can explain the expectations of the item rating task; see Appendix II). Most (but not all) participants who watched the decontextualizing and spoofing videos were able to correctly identify these techniques (72% and 70% respectively). Specifically, participants found the word "decontextualizing" difficult to recall. After watching the video, most participants were able to clearly describe what the technique means. However, they also found the word "decontextualizing" intimidating and tough to remember because it was unfamiliar and hard to pronounce. Further, several participants felt that the term "scapegoating" was too broad a concept to be labelled a manipulation technique. They felt that in many cases, certain groups are indeed to blame for problems and that scapegoating, therefore, cannot be categorized as a manipulation technique per se. Also, even participants who were somewhat familiar with the techniques presented in the videos said the videos endowed them with a "proper understanding" or "more detailed definition" of the techniques. For example, participants said "I have encountered this technique, but didn't know there was a word for it" or "I have encountered this technique online but now I know how to define it". Finally, most participants (even those from semi-urban and mid-sized cities) preferred being asked questions in English or in a mix of English and Hindi, as opposed to only in Hindi. This is important because all of the items in Studies 1-4 were administered in Hindi (or Marathi or Malayalam), not English.

Discussion

In Study 5, we sought to gain a better understanding of why we only found weak support for our hypotheses in Studies 1-4 and why we did not observe the expected improvement on the item rating tasks after the videos. Our results show that the explanation should likely not be sought in the content or presentation of the videos themselves: many if not all participants found the videos entertaining, useful, and necessary, and moreover displayed an improved understanding of relevant manipulation techniques after watching. This alleviates concerns about any potential "backfire" effects, and shows that the videos are effective at their intended goals, namely 1) to capture people's attention and 2) increase awareness and understanding of various common forms of manipulation.

Instead, we find evidence that the item rating tasks administered post-intervention in Studies 1-4 were seen by many participants as confusing; many indicated that they did not understand what was expected of them, and after an explanation indicated that the items (possibly due to information being lost in translation) were not so clear-cut in their usage of manipulation techniques. This was especially the case for the items from Study 4 (the YouTube field study), because they came in the form of simple sentences without additional contextual information (such as formatting it as a WhatsApp message, as we did in Studies 1-3). This meant that participants were unsure how to apply the knowledge they learned in the videos to the items they were subsequently shown, which may explain the inconsistent findings we observed in Studies 1-4. We test these preliminary findings more formally in Study 6.

Study 6: Adapted RCT

Study 6 is an adapted randomized controlled trial with participants from 8 Indian states (Delhi, Gujarat, Madhya Pradesh, Maharashtra, Punjab, Rajasthan, Uttar Pradesh, and West Bengal). The goal of this study was to test the findings from Study 5 in a randomized controlled setting, and to gain further insight into the lack of hypothesized results in Studies 1-3.

Methodology

We recruited a total of 827 participants through a collaboration partner (Ormax Media; see Study 5), who carried out the survey with guidance from Mythos Labs. As in Study 5, participants were internet users from geographically diverse areas of India. We also asked for participants' age and to what extent they appreciated the prebunking video they watched (only to treatment group participants; see below. The scale ran from 1 (excellent) to 5 (poor)). See Table 6 for a sample overview.

Variable	**Category**	**Count**	**%**
Age	Mean	30.5	
	Median	30	
	SD	7.48	
Gender	Female	392	47.4%
	Male	435	52.6%
State	Delhi	156	18.9%
	Gujarat	97	11.7%
	Madhya Pradesh	36	4.4%
	Maharashtra	261	31.6%
	Punjab	61	7.4%
	Rajasthan	59	7.1%
	Uttar Pradesh	114	13.8%
	West Bengal	43	5.2%
Video appreciation	1 (excellent)	346	56.1%
	2 (very good)	214	34.7%
	3 (good, but not very good)	46	7.5%
	4 (average)	10	1.6%
	5 (poor)	1	0.2%

Table 6. Study 6: Sample overview.

Participants were assigned to either a control condition (*n* = 210) where people were not shown any video (and only answered the questions), or one of five treatment conditions where they watched one of the prebunking videos: decontextualization (*n* = 120), emotional manipulation (*n* = 131), fake cures (*n* = 121), scapegoating (*n* = 125) or spoofing (*n* = 120). Treatment group participants were then asked by the moderator to evaluate three statements as either containing or not containing the manipulation technique of interest ("This is [manipulation technique]" or "this is not [manipulation technique]". Two of these statements contained the manipulation technique, and one did not (i.e., it was a neutral statement). As such, responses were coded as either correct (1) or incorrect (0). Control group participants evaluated all 15 statements, i.e., three statements per video. The panel survey was administered digitally, and responses were collected by Ormax Media.

Our outcome variable (the number of correct responses) is ordinal (i.e., an integer between 0 and 3 for each participant), so we fit a series of cumulative link mixed models using the *clmm* function from the lme4 R package (one model per prebunking video). We modelled condition (treatment - control), age, and gender as fixed effects, and participants and region as random effects.

Results

We find that for both the Spoofing video (*OR* = 2.29, *p* = .004) and Decontextualization video (*OR* = 4.00, *p* < .001), participants in the treatment condition were significantly better than the control group at correctly identifying manipulation. However, this was not the case for Scapegoating (*OR* = 0.97, *p* = .876), Fake cures (*OR* = 1.73, *p* = .177) or Emotional manipulation (*OR* = 1.28, *p* = .366). See Table S10 for the full results.

Discussion

In this study, we incorporated insights from Study 5 and made several important adaptations to the study designs used in Studies 1-3. Most importantly, we 1) reduced the number of items, 2) implemented a binary correct/incorrect response mode, and 3) offered participants the opportunity to ask clarifying questions to a study moderator (who, as mentioned, were not allowed to hint at answers or provide specific information about the videos). Doing so resulted in some notable improvements compared to Studies 1-3: two out of five videos (Spoofing and Decontextualization) now showed significant improvement in manipulation technique detection compared to a control group). For the other three videos, no significant differences between the treatment and control group were observed; this also contrasts with Studies 1-3, in which we regularly found *inverse* results, in the sense that treatment group participants appeared to become significantly worse at identifying manipulation. We further reaffirm a preliminary finding from Study 5, namely that the videos were broadly well-liked by viewers: as Table 6 shows, around 90.8% of study participants found the video to be either "excellent" or "very good", with only one participant rating a video as "poor". While there may be demand effects at play (which we are unable to test as we did not obtain data for the control group for this measure), the results are nonetheless encouraging.

Overall, Study 6 yields preliminary evidence that the videos themselves are broadly understood and appreciated by audiences, and in two cases we found that our adapted methodology showed substantial improvements in manipulation technique detection (which Studies 1-3 were unable to pick up on). However, we did not find significant effects in the hypothesized direction for 3 out of 5 cases. This may be in part due to a low sample size or insufficient power (as we had ~ 120 participants in each treatment condition and both the Fake cures and Emotional manipulation videos trended in the right direction). However it remains a possibility that either the updated testing method was still confusing or too cognitively demanding, or simply that these inoculation videos were unsuccessful at conferring psychological resistance to manipulation.

General Discussion and Conclusion

The goal of this project was to assess the efficacy of five “prebunking” videos aimed at countering common forms of online manipulation in India. Across six studies (four preregistered) with more than 135,000 participants, we find mixed support for our hypotheses. While the findings from the online pilot study (Study 1) were somewhat in line with our hypotheses (with several caveats), this was not the case for field studies conducted in Hindi and Malayalam (Studies 2 and 3), and a large scale study conducted on YouTube (in Hindi, Marathi, and Malayalam; Study 4) showed very small effects which, though meeting the standard threshold for statistical significance, are far removed from larger effects reported in previous studies with highly similar research designs (Roozenbeek, van der Linden et al., 2022). Overall, we find that the videos were very well received by the target audiences (qualitative feedback indicated that people greatly enjoyed the videos and thought that they were useful and fun; Study 5, and changes to the study design based on feedback from focus group participants somewhat improved item evaluation task performance; Study 6), but this did not translate to improved performance on the direct messaging app post item rating task in Studies 1-3, which was our key outcome measure of interest. There are several considerations to take into account.

First, in Studies 2 and 3, we find some evidence (specifically for the Decontextualization and Scapegoating videos) that study participants become worse at correctly identifying both manipulative content and “real news”. While it is difficult to provide an explanation for why this is the case using the data collected from this study, there are several possibilities. First, we suspect that survey participants may not have been used to the type of survey that was administered which may have limited the validity of their responses. Our implementation partners for Studies 2 and 3 (Outline India, a highly trusted survey provider) indicated this may have been the case based on their professional experience in conducting surveys in India. Additionally, qualitative feedback from participants indicated that they found the stimuli rating task to be repetitive and boring which may further explain the found response patterns.

Second, it is possible that watching the video led to an increased scepticism of *all* content; however, we deem this to be unlikely, as 1) qualitative responses (by the survey team as well as those collected through focus groups, see Study 5) indicated that participants broadly enjoyed the videos and believed them to be useful, 2) increased scepticism is not observed in other studies with similar scopes and designs, including in our social media Study 4, and 3) true backfire effects, where interventions actively make things *worse*, are exceedingly rare and more likely to be observed due to methodological/design flaws rather than reflecting a real-world phenomenon (Leder et al., 2024; Haglin, 2017; Swire-Thompson et al., 2020, 2022; Wood & Porter, 2019).

Overall, we therefore consider methodological considerations to be a likely explanation for our findings. Considering the wide variety of limitations inherent to our study design, as well as the difficulties associated with conducting misinformation intervention research in the Global South (Badrinathan & Chauchard, 2023; Blair et al., 2024), it is likely that the methodology used in Studies 2 and 3 (and potentially in Study 4, which used a different method) was simply not equipped to measure to what extent successful learning occurred.

The results from Study 5 showed that focus group participants displayed a good understanding of the videos' purpose and scope, and found them to be both humorous and informative. In addition, they displayed increased awareness of the manipulation techniques that the videos were about, and were able to apply this knowledge in an item rating task after being given additional information about the expectations by focus group moderators. These results were partially replicated in Study 6, where we found improved item rating task performance in two out of five videos. We encourage researchers to test interventions in India and other Global South countries and take the above insights into account when designing their interventions and efficacy assessments. With respect to the literature on prebunking and inoculation, we hope that this study helps further understand its cross-cultural and cross-sectional efficacy.

References

Al-Zaman, S. (2022). A Thematic Analysis of Misinformation in India during the COVID-19 Pandemic. *International Information & Library Review*, *54*(2), 128–138. https://doi.org/10.1080/10572317.2021.1908063

Atlani-Duault, L., Ward, J. K., Roy, M., Morin, C., & Wilson, A. (2020). Tracking online heroisation and blame in epidemics. *The Lancet Public Health*, *5*(3), E137–E138. https://doi.org/10.1016/S2468-2667(20)30033-5

Arun, C. (2019). On WhatsApp, Rumours, and Lynchings. *Economic & Political Weekly*, *54*(6).

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

Badrinathan, S., & Chauchard, S. (2023). Researching and Countering Misinformation in the Global South. *Current Opinion in Psychology*, 101733. https://doi.org/10.1016/j.copsyc.2023.101733

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

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

Basuroy, T. (2022). *Social networks used to access news across India in 2022*. Statista. https://www.statista.com/statistics/1026234/india-social-networks-used-to-access-news/

Blair, R. A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., & Stainfield, C. J. (2024). Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology*, *55*, 101732. https://doi.org/10.1016/j.copsyc.2023.101732

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Capewell, G., Maertens, R., Remshard, M., Compton, J., Lewandowsky, S., van der Linden, S., & Roozenbeek, J. (2024). Misinformation interventions decay rapidly without an immediate post-test. *Journal of Applied Social Psychology, 54*(8), 441–454. https://doi.org/10.1111/jasp.13049

Carrasco-Farré, C. (2022). The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, *9*(162). https://doi.org/10.1057/s41599-022-01174-9

Coan, T. G., Boussalis, C., Cook, J., & Nanko, M. O. (2021). Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, *11*(22320). https://doi.org/https://doi.org/10.1038/s41598-021-01714-4

Compton, J. (2013). Inoculation Theory. In J. P. Dillard & L. Shen (Eds.), *The SAGE Handbook of Persuasion: Developments in Theory and Practice* (2nd ed., pp. 220–236). SAGE Publications, Inc. https://doi.org/10.4135/9781452218410

Cook, J., Ecker, U. K. H., Trecek-King, M., Schade, G., Jeffers-Tracy, K., Fessmann, J., Kim, S. C., Kinkead, D., Orr, M., Vraga, E. K., Roberts, K., & McDowell, J. (2022). The Cranky Uncle game—Combining humor and gamification to build student resilience against climate misinformation. *Environmental Education Research*. https://doi.org/10.1080/13504622.2022.2085671

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

Ecker, U. K. H., Roozenbeek, J., van der Linden, S., Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Misinformation poses a bigger threat to democracy than you might think. *Nature, 630*(8015), 29–32. https://doi.org/10.1038/d41586-024-01587-3

Haglin, K. (2017). The limitations of the backfire effect. *Research & Politics*, *4*(3). https://doi.org/10.1177/2053168017716547

Harrop, I., Roozenbeek, J., Madsen, J. K., & van der Linden, S. (2023). Inoculation Can Reduce the Perceived Reliability of Polarizing Social Media Content. *International Journal of Communication*, *16*, 1–24.

Hughes, B., Braddock, K., Miller-Idriss, C., Goldberg, B., Criezis, M., Dashtgard, P., & White, K. (2021). Inoculating against Persuasion by Scientific Racism Propaganda: The Moderating Roles of Propaganda Form and Subtlety. *PsyArxiv Preprints*. https://doi.org/10.31235/osf.io/ecqn4

Ingram, D. (2020, October 26). Twitter launches “pre-bunks” to get ahead of voting misinformation. *NBC News*.

Jigsaw. (2023, October 25). *Prebunking to Build Defenses Against Online Manipulation Tactics in Germany*. Medium. https://medium.com/jigsaw/prebunking-to-build-defenses-against-online-manipulation-tactics-in-germany-a1dbfbc67a1a

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychological Science in the Public Interest*, *21*(3), 103–156. https://doi.org/10.1177/1529100620946707

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Maertens, R., Panizza, F., Pennycook, G., Rand, D. J., Rathje, S., Reifler, J., … Wineburg, S. (2024). Toolbox of Interventions Against Online Misinformation and Manipulation. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-01881-0

Leder, J., Schellinger, L. V., Maertens, R., Chryst, B., van der Linden, S., & Roozenbeek, J. (2024). Feedback Exercises Boost Discernment of Misinformation for Gamified Inoculation Interventions. *Journal of Experimental Psychology: General*, *153*(8), 2068–2087. https://doi.org/10.1037/xge0001603

Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X.-D. (2023). Psychological Inoculation for Credibility Assessment, Sharing Intention, and Discernment of Misinformation: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *25*, e49255. https://doi.org/10.2196/49255

Maertens, R., Roozenbeek, J., Simons, J., Lewandowsky, S., Maturo, V., Goldberg, B., Xu, R., & van der Linden, S. (2024). Psychological Booster Shots Targeting Memory Increase Long-Term Resistance Against Misinformation. *Nature Communications*.

McGuire, W. J., & Papageorgis, D. (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *Journal of Abnormal and Social Psychology*, *62*(2), 327–337.

Pasquetto, I. v, 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/

Patil, I. (2021). Visualizations with statistical details: The "ggstatsplot" approach. *Journal of Open Source Software*, *6*(61), 3167. https://doi.org/10.21105/joss.03167

Pfau, M., Van Bockern, S., & Kang, J. G. (1992). Use of inoculation to promote resistance to smoking initiation among adolescents. *Communications Monographs*, *59*(3), 213–230. https://doi.org/10.1080/03637759209376266

Piltch-Loeb, R., Su, M., Testa, M., Goldberg, B., Braddock, K., Miller-Idriss, C., Maturo, V., & Savoia, E. (2022). Testing the Efficacy of Attitudinal Inoculation Videos to Enhance COVID-19 Vaccine Acceptance: A Quasi-Experimental Intervention Trial. *JMIR Public Health and Surveillance*, *8*(6). https://doi.org/10.2196/34615

Roozenbeek, J., Remshard, M., & Kyrychenko, Y. (2024). Beyond the Headlines: On the Efficacy and Effectiveness of Misinformation Interventions. *Advances in Psychology*, *2*, e24569. https://doi.org/10.56296/aip00019

Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions. *European Psychologist*, *28*(3), 189–205. https://doi.org/10.1027/1016-9040/a000492

Roozenbeek, J., Traberg, C. S., & van der Linden, S. (2022). Technique-based inoculation against real-world misinformation. *Royal Society Open Science*, *9*(211719). https://doi.org/10.1098/rsos.211719

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Humanities and Social Sciences Communications*, *5*(65), 1–10. https://doi.org/10.1057/s41599-019-0279-9

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-2020-47

Roozenbeek, J., & van der Linden, S. (2024). *The Psychology of Misinformation*. Cambridge University Press.

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34). https://doi.org/10.1126/sciadv.abo6254

Simchon, A., Brady, W. J., & Van Bavel, J. J. (2021). Troll and Divide: The Language of Online Polarization. *PsyArxiv Preprints*. https://doi.org/10.31234/osf.io/xjd64

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, *9*(3), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006

Swire-Thompson, B., Miklaucic, N., Wihbey, J., Lazer, D., & DeGutis, J. (2022). Backfire effects after correcting misinformation are strongly associated with reliability. *Journal of Experimental Psychology: General*, *151*(7), 1655–1665. https://doi.org/10.1037/xge0001131

The Hindu. (2019). *India becomes YouTube's largest and fastest growing market*. Www.Thehindu.Com. https://www.thehindu.com/business/india-becomes-youtubes-largest-and-fastest-growing-market/article26785428.ece

Traberg, C. S., 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*(1), 136–151. https://doi.org/10.1177/00027162221087936

Traberg, C. S., & van der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personality and Individual Differences*, *185*(111269). https://doi.org/10.1016/j.paid.2021.111269

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political Psychology in the Digital (mis)Information age: A Model of News Belief and Sharing. *Social Issues and Policy Review*, *15*(1), 84–113. https://doi.org/10.1111/sipr.12077

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 van der Linden, S., & Roozenbeek, J. (2020). Psychological inoculation against fake news. In R. Greifeneder, M. Jaffé, E. Newman, & N. Schwarz (Eds.), *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation* (pp. 147–170). Psychology Press. https://doi.org/10.4324/9780429295379-11

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

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/10.1080/13504630.2020.1782730

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wood, T., & Porter, E. (2019). The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior, 41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y