


# Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation

Big Data & Society  
 January–June: 1–18  
 © The Author(s) 2021  
 Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
 DOI: 10.1177/20539517211013868  
[journals.sagepub.com/home/bds](https://journals.sagepub.com/home/bds)  


Melisa Basol<sup>1</sup> , Jon Roozenbeek<sup>1</sup> , Manon Berriche<sup>2</sup> ,  
 Fatih Uenal<sup>1</sup> , William P. McClanahan<sup>1</sup>  and  
 Sander van der Linden<sup>1</sup> 

## Abstract

Misinformation about the novel coronavirus (COVID-19) is a pressing societal challenge. Across two studies, one preregistered ( $n_1 = 1771$  and  $n_2 = 1777$ ), we assess the efficacy of two ‘prebunking’ interventions aimed at improving people’s ability to spot manipulation techniques commonly used in COVID-19 misinformation across three different languages (English, French and German). We find that *Go Virall*, a novel five-minute browser game, (a) increases the perceived manipulateness of misinformation about COVID-19, (b) improves people’s attitudinal certainty (confidence) in their ability to spot misinformation and (c) reduces self-reported willingness to share misinformation with others. The first two effects remain significant for at least one week after gameplay. We also find that reading real-world infographics from UNESCO improves people’s ability and confidence in spotting COVID-19 misinformation (albeit with descriptively smaller effect sizes than the game). Limitations and implications for fake news interventions are discussed.

## Keywords

Fake news, misinformation, inoculation theory, prebunking, COVID-19, gamification

This article is a part of special theme on Studying the COVID-19 Infodemic at Scale. To see a full list of all articles in this special theme, please click here: <https://journals.sagepub.com/page/bds/collections/studyinginfodemicatscale>

## Introduction

The SARS-CoV-2 (COVID-19) pandemic is a pressing global health crisis, the mitigation of which relies in part on non-pharmaceutical interventions that leverage insights from the social and behavioural sciences (Van Bavel et al., 2020; Van der Linden et al., 2020). Misinformation about the disease has spread widely on social media, ranging from fake ‘remedies and cures’, such as eating garlic or injecting bleach, to elaborate conspiracy theories behind the cause of COVID-19 (BBC News, 2020). In response, the World Health Organization (WHO) has warned of an ‘infodemic’ (Zarocostas, 2020), with some claiming that the prevalence of misleading information around the virus might be ‘the most contagious thing about it’ (Kucharski, 2020). Susceptibility to misinformation about

COVID-19 relates to a variety of negative outcomes (Enders et al., 2020; Roozenbeek et al., 2020b), such as affecting people’s willingness to comply with evidence-based health regulations (Imhoff and Lamberty, 2020), support for violence (Jolley and Paterson, 2020) and vaccine uptake intentions around

<sup>1</sup>Department of Psychology, School of the Biological Sciences, University of Cambridge, Cambridge, UK

<sup>2</sup>Sciences Po médialab, Center for Research and Interdisciplinarity, Paris, France

### Corresponding author:

Sander van der Linden, Department of Psychology, University of Cambridge, Downing Street, CB2 3EB Cambridge, UK.  
 Email: [sander.vanderlinden@psychol.cam.ac.uk](mailto:sander.vanderlinden@psychol.cam.ac.uk)



the world (Loomba et al., 2021; Roozenbeek et al., 2020b).

To add to the problem, information about COVID-19 is not always easily classified as either true or false. Considering the continuously developing scientific understanding of the virus, information around it can range between various degrees of unverified or disproven, making a clear classification of what counts as ‘misinformation’ difficult (Vraga and Bode, 2020). Nevertheless, in a Pew survey, almost half of the sample (48%) reported having been exposed to falsehoods about the virus, the majority of whom claimed to see misleading information on a daily basis (Schaeffer, 2020). Another study found that 85% of U.S. participants believed more than one false or misleading statement about COVID-19 (Miller, 2020). Frequent exposure to misinformation is particularly dangerous as repetition increases reliance on false information (Fazio et al., 2015). As the supply of evidence-based interventions remains low (Agly et al., 2020; Pennycook et al., 2020), it is critical to explore how the spread of misinformation around COVID-19 may be mitigated.

### Theoretical background: Prebunking and inoculation theory

Preemptively debunking (‘prebunking’) misinformation is regarded as a promising step towards building attitudinal resistance against misinformation. Prebunking is a key component of inoculation theory, often regarded as the ‘grandfather theory of persuasion’ (Eagly and Chaiken, 1993: 561). Psychological inoculation is based on a biological analogy of the immunisation process (McGuire, 1964). Similar to how exposure to a weakened dosage of a pathogen triggers the generation of protective antibodies, inoculation theory posits that a weakened persuasive argument will elicit motivation to equip oneself with protective arguments against it (McGuire and Papageorgis, 1961). Both processes thus rely on the assumption that exposure to a weakened pathogen triggers an immunity-bolstering response. In the psychological inoculation literature, the inoculation process commonly consists of two elements: (1) a forewarning – which elicits threat and motivation to defend one’s attitudes – and (2) a pre-emptive refutation (or ‘prebunk’) of the persuasive arguments (Compton, 2013; Compton and Pfau, 2005). Research has demonstrated the robustness and efficacy of psychological inoculations in conferring resistance across a multitude of topics (for reviews and meta-analyses, see Banas and Rains, 2010; Compton et al., 2021; Lewandowsky and van der Linden, 2021).

Yet, the inoculation analogy is meant to be ‘more instructive than prescriptive’, and several important open questions about the boundary conditions of inoculation theory remain (Compton, 2013: 233). Specifically, recent inoculation research has seen three key innovations. First, researchers have begun to distinguish between *prophylactic* and *therapeutic* inoculation approaches (Compton, 2019). Inoculation can be fully preemptive (prophylactic) when people have not yet been exposed to misinformation. In contrast, therapeutic inoculation treatments – like therapeutic medical vaccines – are administered to those who have already had some potential prior exposure to misinformation as they can still boost immunity and limit further spread among the ‘already afflicted’ (Compton, 2019; Wood et al., 2012). In practice, this distinction matters little as people are inoculated regardless of prior exposure, but theoretically it helps to elucidate the conditions under which inoculation can be effective (Compton, 2019; Compton et al., 2021; van der Linden and Roozenbeek, 2020).

Second, traditional psychological inoculations often target specific false or misleading arguments, for example about climate change (Cook et al., 2017; van der Linden et al., 2017b) or vaccinations (Jolley and Douglas, 2017). This issue-based approach makes it difficult to scale inoculation interventions. To help scale the approach further, recent work has shifted away from inoculating against individual arguments towards conferring resistance against the *techniques* that underlie many instances of misinformation (Cook et al., 2017; Roozenbeek and van der Linden, 2018; van der Linden et al., 2021).

Third, inoculation research has recently explored the benefits of ‘active’ versus ‘passive’ inoculations (Roozenbeek and van der Linden, 2018). Traditional inoculation research has predominantly required participants to passively read a short text as part of the inoculation treatment (Banas and Rains, 2010; Compton, 2013). More recent work has produced interventions that simulate social media environments in which people are prompted to make decisions proactively, thus generating their own ‘mental antibodies’. An example of such an active inoculation intervention is the award-winning ‘fake news’ game *Bad News* (Basol et al., 2020; Maertens et al., 2020; Roozenbeek and van der Linden, 2019; Roozenbeek et al., 2020a), in which players build psychological resistance against six common misinformation techniques. Yet, little is currently known about the differences between active and passive inoculation treatments (Banas and Rains; Compton et al., 2021), their persistence over time (Maertens et al., 2020) and their effect on attitudinal certainty (Basol et al., 2020).

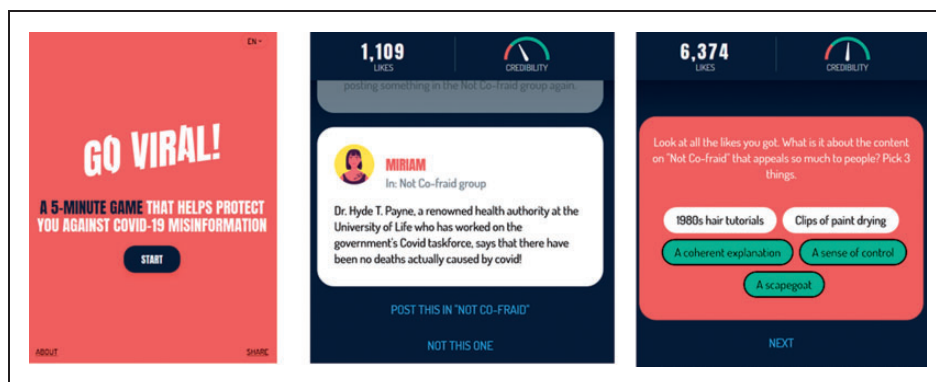
In addition, several other important gaps in our understanding of inoculation theory remain: relatively little attention is paid to how psychological inoculations affect how likely people are to share misinformation (but see Roozenbeek and van der Linden, 2020). Reducing the spread of misinformation and sharing the ‘vaccine’ are important elements in achieving psychological ‘herd immunity’ (Compton and Pfau, 2009; van der Linden et al., 2017a), particularly within the context of the COVID-19 pandemic. Furthermore, how inoculation interventions affect attitudinal certainty remains an open question. Research suggests that the more certain individuals are of their attitudes, the more likely an inoculation is to guide behaviour (Rucker and Petty, 2004), help resist persuasion (Tormala and Petty, 2002), and persist over time (Tormala, 2016). In the context of misinformation, it is therefore important to explore if inoculation interventions not only reduce susceptibility to misinformation, but also to what extent people become more confident in their ability to spot it and whether they are less likely to share it with others.

## The present research

In this study, we address these gaps in the inoculation literature within the context of COVID-19 misinformation. To do so, we test two technique-based prebunking interventions aimed at improving people’s ability to spot misinformation about COVID-19 alongside each other. The first intervention is *Go Viral!*, a novel and freely available five-minute choice-based browser game similar in design to other ‘fake news’ games such as *Bad News* (Roozenbeek and van der Linden, 2019) and *Harmony Square* (Roozenbeek and van der Linden, 2020). We created *Go Viral!* ([www.goviralgame.com](http://www.goviralgame.com)) in collaboration with the UK Cabinet Office and DROG with support from the WHO and the United Nations’ Verified Campaign to expose three manipulation techniques commonly used in COVID-19

misinformation: fearmongering, using fake experts, and spreading conspiracy theories (World Health Organization, 2020b; Zarocostas, 2020). *Go Viral!* is available in three languages (English, French, and German), is listed by the WHO as an anti-misinformation resource<sup>1</sup> and has been played approximately 300,000 times since its launch in October 2020. *Go Viral!* functions as an active inoculation against future manipulation attempts by pre-emptively warning and exposing people to weakened doses of COVID-19 misinformation and letting them generate their own psychological ‘antibodies’ (van der Linden et al., 2020).

In the game, players start out by browsing their (fictitious) social media feed and are slowly lured into an echo chamber where misinformation and outrage-evoking content about COVID-19 are common (these scenarios are aimed at eliciting threat and motivation). Across three scenarios, players are encouraged to gain ‘likes’ and ‘credibility points’ while learning about three common manipulation techniques. In the first scenario, ‘The Fearmongerer’, players create a social media post by using emotionally evocative language and watch it go viral. The use of moral-emotional language is known to enhance the virality of social media content (Acerbi, 2019; Berriche and Altay, 2020; Brady et al., 2017). They are then invited to join *Not Co-Fraid*, a group of online ‘truth tellers’. In the second scenario, ‘My Imaginary Expert’, players start sharing content in the group as *Not Co-Fraid*’s latest member. Their low credibility, however, prompts them to back up their claims by using fake experts, such as Dr Hyde T. Paine from the ‘University of Life’. By giving *Not Co-Fraid* group members the illusion that their content is endorsed by experts, players gain popularity, and are eventually asked to become a *Not Co-Fraid* moderator. This scenario relies on impersonation and the fake expert technique, both of which are commonly used in online misinformation (Cook et al., 2017; Roozenbeek and van der Linden, 2019). In the final scenario, ‘Master of Puppets’, players create their



**Figure 1.** *Go Viral!* landing page (left) and game environment (middle and right).



own COVID-19 conspiracy theory. They first pick a target (e.g. a large NGO, the government or one Bob from New York), accuse it of shady practices and connect the dots, resulting in nationwide protests. Conspiracy theories have featured heavily around COVID-19 (van der Linden et al., 2020) and have been linked to violent intentions (Jolley and Paterson, 2020) and reduced willingness to comply with health guidelines (Roozenbeek et al., 2020b). Figure 1 shows the *Go Viral!* landing page and game environment.

The second prebunking intervention consists of a series of infographics about COVID-19 misinformation. As part of its #ThinkBeforeSharing prebunking campaign, UNESCO, with input from inoculation researchers, created a social media package of images that explain how COVID-19 misinformation is created and spreads (UNESCO, 2020). Figure 2 shows several examples.

In this study, we leverage the public availability of both interventions to test a number of key hypotheses pertaining to prebunking and inoculation theory as a way to reduce susceptibility to misinformation. First, this study advances the literature by testing prebunking interventions in the context of COVID-19 misinformation. Second, to date, no published research has assessed different types of anti-misinformation inoculation or prebunking interventions alongside each other. Crucially, some findings suggest that so-called ‘active’ inoculation (e.g. in form of a game) confers attitudinal resistance more effectively than ‘passive’ inoculation (i.e. through reading, see Banas and Rains, 2010; McGuire and Papageorgis, 1961; Roozenbeek and van der Linden, 2018). This study is the first to address this question in the context of misinformation. Third, this study is one of the first to explore how prebunking interventions affect self-reported measures of behaviour; in this case, people’s willingness to share misinformation with others

(Roozenbeek and van der Linden, 2020). Fourth, we build on the existing literature on attitudinal certainty by exploring how such interventions affect people’s confidence in their ability to spot misinformation. Fifth, following recommendations to maximise the generalisability of interventions effects (O’Keefe, 2015), we also make use of the public availability of both *Go Viral!* and the UNESCO infographics in English, French and German to assess the effectiveness of prebunking interventions in different cultural and linguistic settings (Roozenbeek et al., 2020). Finally, given the decay of resistance to persuasion effects (Maertens et al., 2020), we evaluate the long-term effectiveness of both interventions after a one-week follow-up.

We address the above questions in two high-powered and large-sample studies. Our Open Science Framework (OSF) page contains all the necessary information needed to replicate our findings and methods, including our datasets, Qualtrics surveys, the full list of items (social media posts), preregistrations, supplementary tables, figures and analyses, and our analysis and visualisation scripts: <https://osf.io/mbqwj/>. Both studies were approved by the Cambridge Psychology Research Ethics Committee (PRE.2020.035).

## Study 1

### Method

In study 1, we implemented a voluntary pre–post survey within the *Go Viral!* game, following the within-subject paradigm developed by Roozenbeek and van der Linden (2019), which is relatively unaffected by testing effects (Roozenbeek et al., 2020a). At the start of the game, players were asked to participate in a scientific study. Consenting participants were shown three misinformation and three real news social



Figure 2. UNESCO infographics.

media posts (in the form of Tweets) relating to COVID-19, and asked to rate the manipulateness of each post on a 1–7 Likert scale (1 being ‘not at all’ and 7 being ‘very’, following Saleh et al., 2021). After completing the game, players were asked to participate in the second part of the study. Upon agreeing to do so, they were again asked to rate the manipulateness of the same social media posts that they saw in the pre-test, and presented with a series of demographic questions: age group, gender, education, political ideology (1 being ‘very left-wing’ and 7 being ‘very right-wing’) and geographic region. Participants received no financial compensation.

The three misinformation posts each make use of a manipulation technique players learn about in the game (using moral-emotional language, using fake experts and conspiratorial reasoning), and were taken from fact-checking websites such as FullFact and the WHO’s COVID-19 Mythbusters page (World Health Organization, 2020a). The three real posts are Tweets about COVID-19, taken from the Twitter accounts of reputable news sources (BBC News, AP, Reuters). To avoid potential source confounds, all source information (for real and fake posts) was blacked out so that assessments were restricted to wording and language use. All social media posts used in this study can be found in the ‘items’ folder on our OSF page: <https://osf.io/mbqwj/>. Figure 3 shows the survey in the in-game environment.

This study design allows us to test the following hypotheses about the effectiveness of *Go Viral!* as a way to improve people’s ability to spot misinformation about COVID-19:

- H<sub>1</sub>:** People who play *Go Viral!* will rate misinformation about COVID-19 as significantly more manipulative after gameplay.
- H<sub>2</sub>:** People who play *Go Viral!* will be able to distinguish real news and misinformation about COVID-19 more accurately after gameplay.

In addition, we test the following null hypothesis:

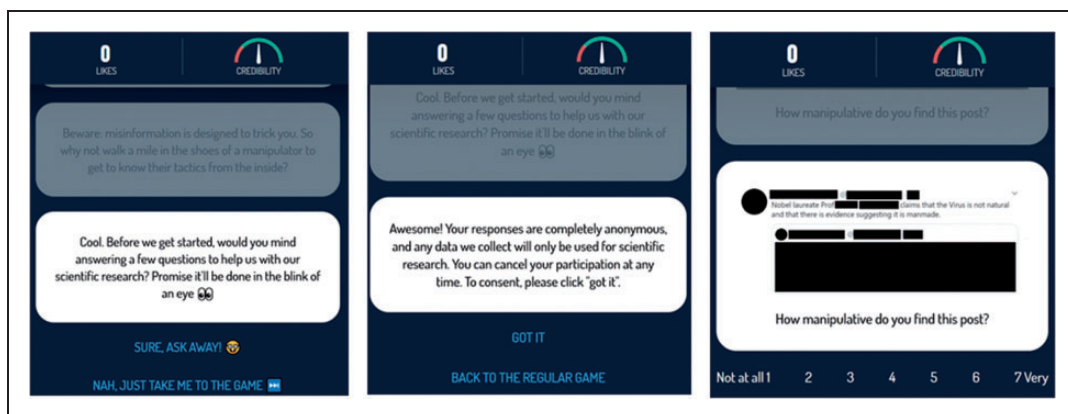
- H<sub>0,realnews</sub>:** People who play *Go Viral!* will not rate real news as significantly more manipulative after gameplay.

## Sample

Between 27 October and 26 November 2020, a total of 2634 complete pre–post survey responses were collected within the *Go Viral!* game environment, out of 14,755 people who completed the game in this time period (a response rate of 17.9%). As per our ethics approval, we excluded 863 underaged participants, leaving a total sample of  $N=1771$ ; 52.9% of our sample identified as male (43.0% female, 1.8% other, 2.3% prefer not to say); 53.6% indicated being between 18 and 34 years of age, and 36.3% reported having a university bachelor’s degree. Our sample also skewed politically left ( $M=3.07$ ,  $SD=1.24$ ). Finally, most study participants were from Europe (59.3%) and North America (22.7%). See Table S1 for the full sample composition.

## Results

To test hypothesis **H<sub>1</sub>**, we conducted a paired-samples *t*-test on the averaged pre- and post-manipulateness scores for the three misinformation items.<sup>2</sup> We find that participants rate misinformation about COVID-19 as significantly more manipulative after playing *Go Viral!* ( $M_{misinformation.pre}=5.61$ ,  $M_{misinformation.post}=6.07$ ,  $M_{diff}=0.46$ , 95% CI (0.419–0.502),  $t(2,1770)=21.88$ ,  $p<0.001$ ,  $d=0.52$ , 95% CI (0.470–0.569)). We find the same result for the individual emotion item ( $M_{emotion.pre}=5.40$ ,  $M_{emotion.post}=5.83$ ,  $M_{diff}=0.43$ , 95% CI (0.367–0.495),  $t(2,1770)=13.21$ ,  $p<0.001$ ,  $d=0.31$ , 95% CI (0.266–0.362)), the fake expert item ( $M_{fakeexpert.pre}=5.45$ ,  $M_{fakeexpert.post}=6.15$ ,  $M_{diff}=0.70$ , 95% CI (0.634–0.763),  $t(2,1770)=21.11$ ,  $p<0.001$ ,  $d=0.50$ , 95% CI (0.452–0.551)), and the conspiracy item ( $M_{conspiracy.pre}=5.98$ ,



**Figure 3.** In-game survey screenshots: start of the survey (left), consent form (middle) and a social media post (right).

$M_{conspiracy.post} = 6.23$ ,  $M_{diff} = 0.25$ , 95% CI (0.196–0.308),  $t(2,1770) = 8.77$ ,  $p < 0.001$ ,  $d = 0.21$ , 95% CI (0.161–0.255)). These results support hypothesis **H<sub>1</sub>**.

To test hypothesis **H<sub>2</sub>**, we conducted a paired samples *t*-test on the pre- and post-gameplay difference for the difference in manipulateness scores between misinformation and real news (i.e. the level of ‘veracity discernment’, or misinformation manipulateness minus real news manipulateness), before and after playing. Doing so gives a significant post-gameplay increase in veracity discernment, showing that *Go Viral!* players are better able to distinguish real news and misinformation about COVID-19 after playing, in support of hypothesis **H<sub>2</sub>** ( $M_{discernment.pre} = 2.98$ ,  $M_{discernment.post} = 3.41$ ,  $M_{diff} = 0.43$ , 95% CI (0.374–0.487),  $t(2,1770) = 14.94$ ,  $p < 0.001$ ,  $d = 0.36$ , 95% CI (0.307–0.403)).

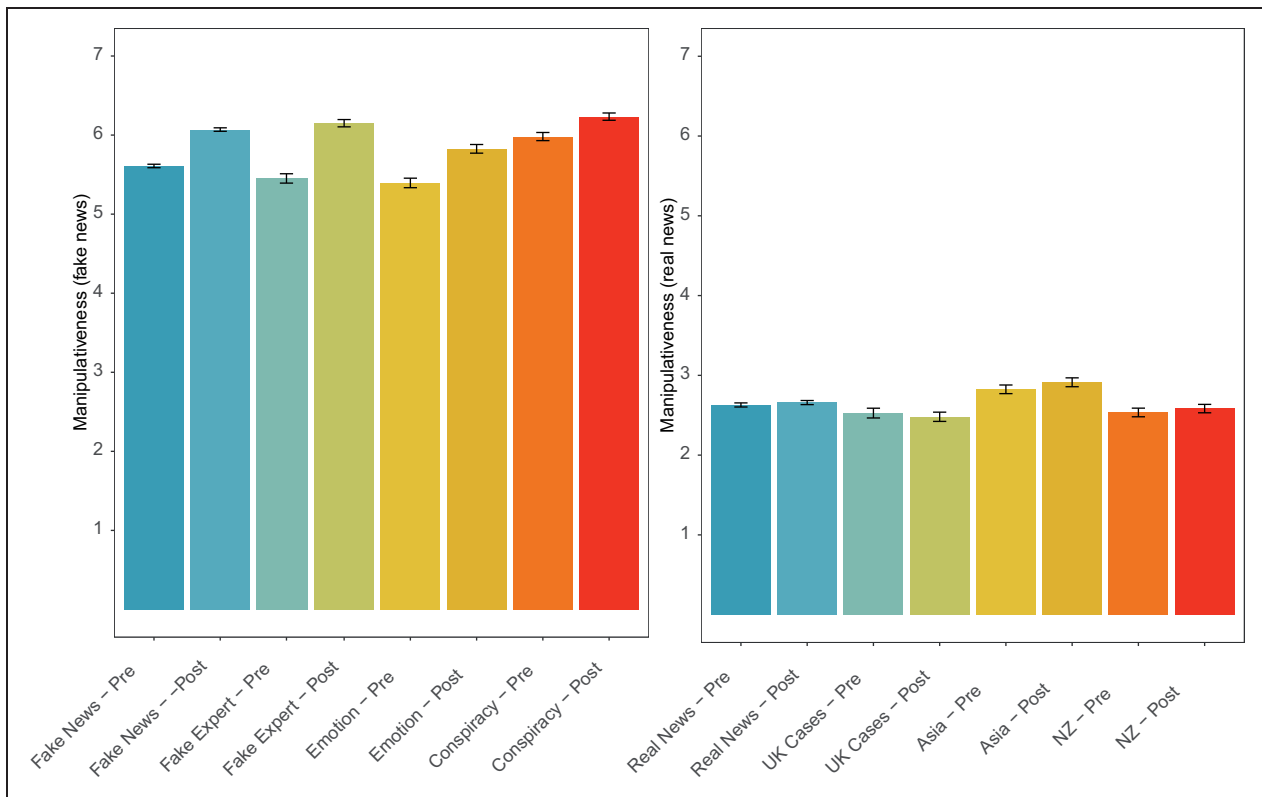
For real news, we find no significant difference in overall pre–post manipulateness scores ( $M_{realnews.pre} = 2.63$ ,  $M_{realnews.post} = 2.66$ ,  $M_{diff} = 0.03$ , 95% CI (–0.0180 to 0.0782),  $t(2,1770) = 1.23$ ,  $p = 0.22$ ,  $d = 0.03$ , 95% CI (–0.0174 to 0.0758)). Furthermore, we find no significant pre–post differences for two out of three real news items either (both  $ps > 0.13$ ), and a small but significant increase in the perceived manipulateness of one real item ( $M_{asia.pre} = 2.83$ ,  $M_{asia.post} = 2.91$ ,  $M_{diff} = 0.09$ , 95% CI (0.0192–

0.157),  $t(2,1770) = 2.51$ ,  $p = 0.012$ ,  $d = 0.06$ , 95% CI (0.013–0.106)). A Bayesian paired samples *t*-test for the averaged real news items gives a Bayes factor of  $BF_{10} = 0.057$  (error % = 0.046), indicating strong support for the null hypothesis **H<sub>0,realnews</sub>** (see van Doorn et al., 2020).<sup>3</sup> Figure 4 shows the results for misinformation and real news in a bar graph.

Finally, to check for covariate effects, we conducted a linear regression with the difference in pre–post veracity discernment as the dependent variable, and gender, age group, education level, political ideology and being from Europe (as this was the largest single geographic region of origin in our sample) as covariates. We find no significant effects (all  $ps > 0.082$ ), except for political ideology ( $p = 0.006$ ), so that identifying as left-wing is associated with a higher post–pre inoculation effect in terms of veracity discernment than people who identify as right-wing. See Table S3.

## Discussion

In a large-sample in-game survey experiment, we showed that people who play *Go Viral!*, irrespective of their demographic background (aside from political ideology), found misinformation about COVID-19 significantly more manipulative after playing than before, whereas their assessment of real news did not change in



**Figure 4.** Bar graph of the perceived manipulateness of fake news (left) and real news (right), averaged and per individual item. Error bars show 95% confidence intervals.

a meaningful sense. The effect sizes are in line with previous studies that have used similar designs (Roozenbeek and van der Linden, 2019), and are particularly encouraging considering these are within-subjects effects. Although this study allowed us to leverage the popularity of *Go Viral!* to collect survey responses ‘in the wild’, it does not include a comparison with other interventions aimed at reducing susceptibility to COVID-19 misinformation. Furthermore, the absence of a randomised control group allows for limited causal inference, and we only ran the survey in one language (English). Finally, to avoid overburdening game players, we only included a total of six items and one outcome measure (manipulativeness). We address these issues in Study 2.

## Study 2

### Method

We conducted a preregistered randomised controlled trial on Prolific Academic with three conditions (an active condition, a ‘passive’ Infographics condition, and a control condition), across three languages: English (using a national sample of the United Kingdom), French and German.<sup>4</sup>

The active (inoculation) condition involved playing *Go Viral!* The Infographics condition involved reading through the UNESCO infographics.<sup>5</sup> The control condition involved attentively playing *Tetris* for a mandatory minimum of five minutes, approximately the same amount of time it takes to complete *Go Viral!* We chose *Tetris* for several reasons: (1) it has been used as a control condition in previous studies on inoculation games (Basol et al., 2020; Maertens et al., 2020; Roozenbeek and van der Linden, 2020); (2) it is in the public domain; and (3) it is a simple game with a flat learning curve.

To begin with, participants performed an item-rating task, where they were randomly shown nine real and nine misinformation social media posts (in the form of tweets, in English, French or German) about COVID-19. Six of these 18 items were the same as those used in Study 1, the other 12 were selected using the same procedure as described in Study 1.<sup>6</sup> In total, participants thus saw nine real news posts (not containing misinformation) and nine misinformation posts (three per manipulation technique: fearmongering, fake experts and conspiracy). As in Study 1, all source information was blacked out to avoid source confounds (Roozenbeek and van der Linden, 2020). We included three main preregistered outcome measures. For each post, participants rated the following statements on a 1–7 scale (1 being ‘strongly disagree’ and 7 being ‘strongly agree’): (1) this post is manipulative (Saleh et al., 2021); (2) I am confident in my assessment of this post’s manipulativenness (attitudinal certainty; Basol et al., 2020); (3) I would share this post with people in my network (Roozenbeek and van der Linden, 2020). See Figure 5.

After completing this item rating task, participants were randomly assigned to one of the treatment conditions (active inoculation or Infographics) or the control condition (1:1:1). Both treatment conditions were followed by manipulation checks to ensure that participants paid sufficient attention. As preregistered, low-effort responses, i.e. participants who gave exactly the same answers for all 18 social media posts in the pre-intervention item rating task and participants who failed one (in the Infographics condition) or two (in the *Go Viral!* condition) attention checks were excluded and resampled.<sup>7</sup> Next, participants were given two tasks in a random order<sup>8</sup>: (1) a set of questions about perceived motivational and apprehensive threat adjusted to the context of misinformation about COVID-19 (Miller et al., 2013; Richards and Banas, 2018)<sup>9</sup>; and (2)

Figure 5 displays two examples of social media posts used in the item rating task. The left panel shows a manipulative post, and the right panel shows a real post. Both posts are followed by a 7-point Likert scale for three rating items: 'This post is manipulative', 'I am confident in my judgment about this post's manipulativenness', and 'I would share this post with people in my network'.

**Left Panel (Manipulative Post):**

Did you know that the new Covid-19 vaccine will be the first of its kind EVER? [redacted] inc. (a pharmaceutical company) said on its website that it will be an mRNA vaccine, which can literally alter your DNA. It will wrap itself into your system.

Strongly disagree (1-2) | Neutral (3-4) | Strongly agree (5-7)

This post is manipulative: 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○

I am confident in my judgment about this post's manipulativenness: 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○

I would share this post with people in my network: 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○

**Right Panel (Real Post):**

UK reports 4,044 COVID-19 cases on Monday compared with 5,693 on Sunday

Strongly disagree (1-2) | Neutral (3-4) | Strongly agree (5-7)

This post is manipulative: 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○

I am confident in my judgment about this post's manipulativenness: 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○

I would share this post with people in my network: 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○

**Figure 5.** Examples of a manipulative (left) and real (right) social media post from the item rating task (Study 2).



the same item rating task that participants completed in the pre-test (i.e. the post-test). After these two tasks, participants answered a series of questions: the ‘vigilance’ measure from the Reuters Digital News Report (a measure of the extent to which people are concerned about the accuracy and source reliability of the news that they consume, with responses ranging from ‘never’ to ‘frequently’ on a four-point scale; see Newman *et al.*, 2020); perceived resistance against misinformation (1–7; see Ivanov *et al.*, 2012); motivation to counter-argue against misinformation about COVID-19 (1–7; see Ivanov *et al.*, 2017); people’s willingness to share the game (*Tetris/Go Viral!*) or the UNESCO infographics on social media accounts (1–7) and in real life (1–7); whether people have had COVID-19 (yes/no/unsure/prefer not to say; see Dryhurst *et al.*, 2020); how worried they are about COVID-19 (1–7; see Dryhurst *et al.*, 2020); and whether they would get vaccinated against COVID-19 if a vaccine became available (yes/no; see Roozenbeek *et al.*, 2020b). Finally, participants were asked several standard demographic questions: birth year, gender, education, and political ideology (1 being ‘very left-wing’ and 7 being ‘very right-wing’).

A week later, UK participants who completed the initial study were reinvited to partake in a follow-up,<sup>10</sup> in which they completed the same item rating task (with manipulativeness, confidence and willingness to share as outcome measures) for 12 new, previously unseen social media posts (6 real and 6 misinformation, or 2 misinformation posts per technique learned in *Go Viral!*). The flowchart in Figure 6 shows the study’s design schematically.

For study 2, we tested the following hypotheses (all preregistered except **H<sub>8</sub>**):

- H<sub>3</sub>**: Participants in both the *Go Viral!* and the Infographics treatment conditions will assess the manipulativeness of real and misinformation more accurately than the control condition, in all three languages.
- H<sub>4</sub>**: Participants in both treatment conditions will be more confident in their manipulativeness assessments than the control condition, in all three languages.
- H<sub>5</sub>**: Participants in both treatment conditions will be less willing to share misinformation with others in their network than the control condition, in all three languages.
- H<sub>6</sub>**: Participants in the active inoculation condition (*Go Viral!*) will be more willing to share the treatment (with others) than the Infographics condition, in all three languages.
- H<sub>7a</sub>**: One week after exposure to the intervention, participants in the inoculation conditions will display minimal decay of the inoculation effect (for manipulativeness, confidence, and sharing).
- H<sub>7b</sub>**: One week after exposure to the intervention, participants in the infographics condition will display

significant decay of the inoculation effect (for manipulativeness, confidence, and sharing).

Finally, based on our preregistered exploratory analyses on threat, we hypothesise that:

- H<sub>8</sub>**: Perceived threat about COVID-19 misinformation is significantly higher in the inoculation condition compared to the Infographics and control conditions.

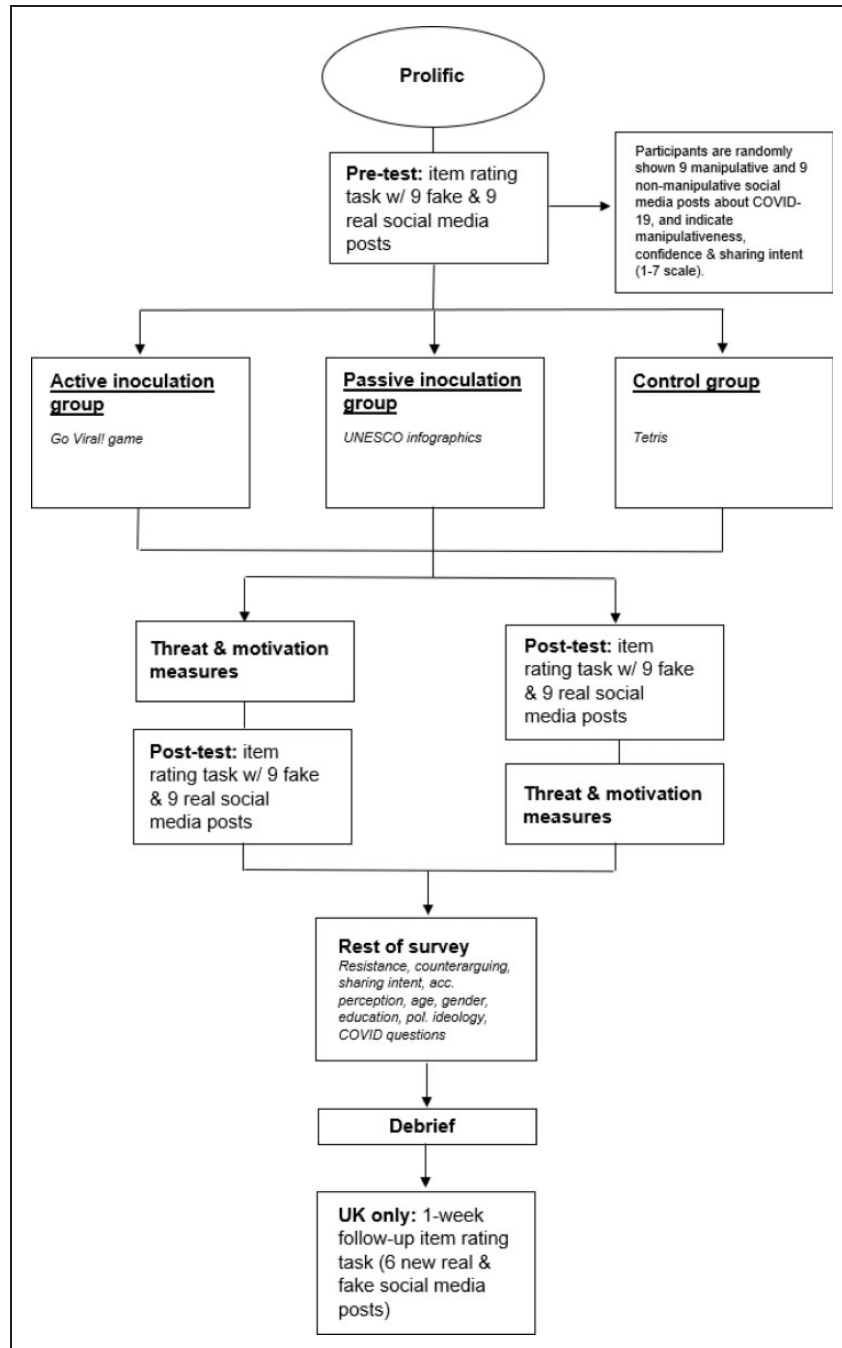
## Sample

Participants were recruited via Prolific Academic (Peer *et al.*, 2017). We first conducted a pilot study ( $n = 231$ ) as a pre-test, in order to validate our item sets.<sup>11</sup> Next, we ran the full study in three different languages: one with a national sample of the UK (in English), one in French and one in German.<sup>12</sup> Participants were paid GBP 1.75 for their participation. UK participants who took part in the follow-up study were paid an additional GBP 0.25. Participants in the Infographics and *Go Viral!* conditions were subjected to one (for Infographics) or two (for *Go Viral!*) attention checks. As per our preregistration, low-effort participants were excluded from the analysis. A priori power analysis using G\*Power with an effect size of  $d = 0.40$ , 95% power, three groups and three measurements (pre – post – follow-up) gives a desired sample size of  $n = 261$  per language to detect a main effect. However, because the effect size is expected to be smaller for the confidence and sharing measures (Roozenbeek and van der Linden, 2020), and in line with our preregistration, we aimed to recruit 900 participants for each language. Unfortunately, due to an unexpected number of participants failing to provide the correct game completion password, the final sample consists of  $n = 710$  valid participants for the UK study,  $n = 610$  for the French study and  $n = 457$  for the German study, for a total of  $N = 1777$ . In total, 606 out of 701 UK participants took part in the one-week follow-up study (86% retention). See Table S4 for the full sample composition by each country.

## Results

We first present the (preregistered) analyses for our three main outcome measures included in the social media posts item rating task (manipulativeness, confidence, and sharing) separately, for both the misinformation and real items, focusing primarily on the *difference* for each outcome measure before and after the intervention between conditions (i.e. hypotheses **H<sub>3</sub>**, **H<sub>4</sub>**, **H<sub>5</sub>**, **H<sub>7a</sub>** and **H<sub>7b</sub>**).<sup>13</sup> Finally, we present the analyses for hypotheses **H<sub>6</sub>** and **H<sub>8</sub>**, which were not part of the item rating task and were only assessed post-intervention (see Figure 6). Our preregistered



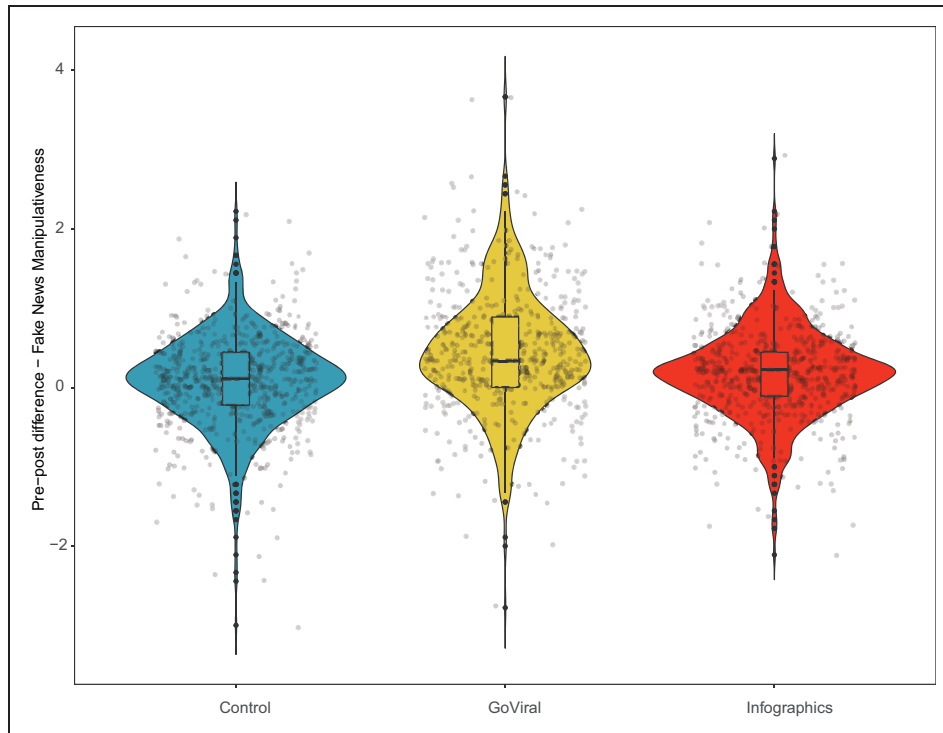


**Figure 6.** Study 2 design flowchart.

robustness checks for these results, including how they differ across covariates (age, gender, education, etc.), as well as the analyses for our vigilance measure (Newman et al., 2020) can be found in the Supplementary Analyses section of the supplement.

**Manipulativenes.** For the pooled sample, a one-way between-subjects ANOVA shows a significant effect of condition (control, *Go Viral!*, Infographics) on the

pre–post intervention difference in the perceived manipulativenes of misinformation about COVID-19 ( $F(2,1774) = 51.69, p < 0.001, \eta^2 = 0.055$ ).<sup>14</sup> A Tukey HSD post-hoc comparison shows that the pre–post difference in perceived manipulativenes for the *Go Viral!* condition was significantly higher than the control condition ( $M = 0.45$  vs  $M = 0.08, M_{diff} = 0.37, 95\% \text{ CI } (0.28\text{--}0.46), p_{tukey} < 0.001, d = 0.56$ ) and the Infographics condition ( $M = 0.45$  vs  $M = 0.18,$



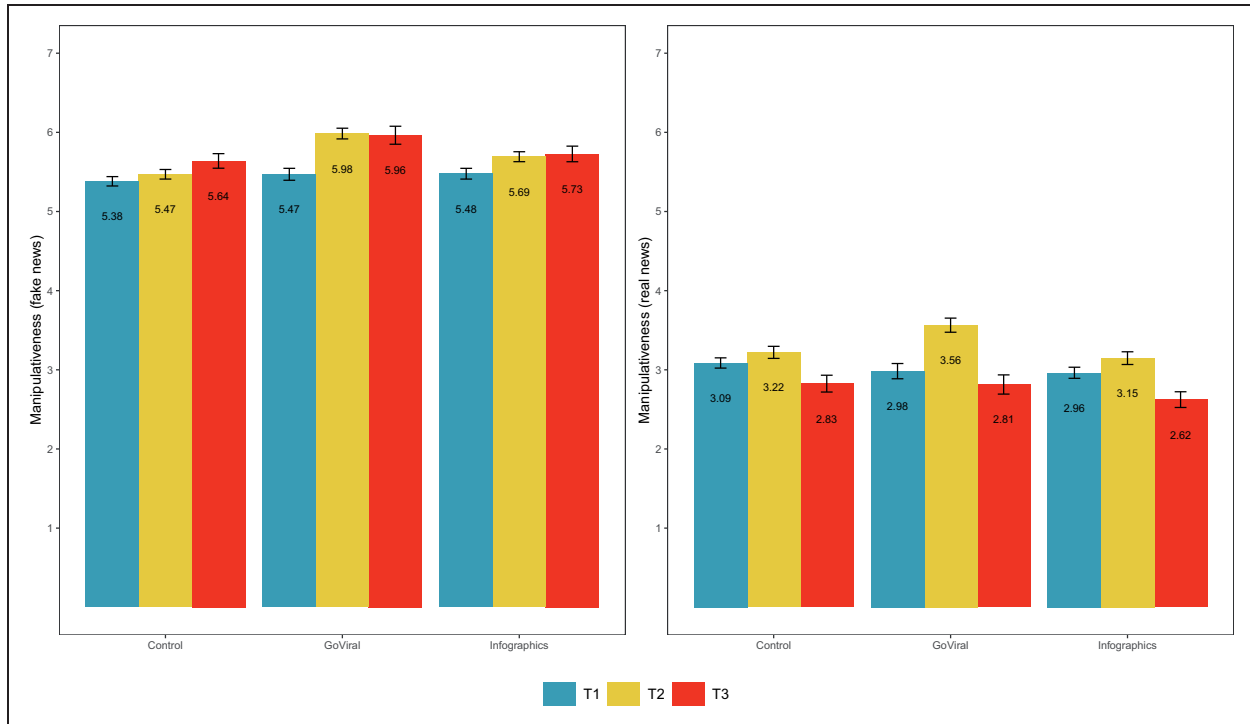
**Figure 7.** Violin plot with jitter of post-pre manipulateness scores of fake news posts (all countries combined).

$M_{diff} = 0.27$ , 95% CI (0.18–0.36),  $p_{tukey} < 0.001$ ,  $d = 0.41$ ). We find a similar effect in the same direction for the Infographics condition compared to the control condition ( $M = 0.18$  vs  $M = 0.08$ ,  $M_{diff} = 0.10$ , 95% CI (0.02–0.19),  $p_{tukey} = 0.015$ ,  $d = 0.17$ ), indicating that both playing the *Go Viral!* game and reading through the UNESCO infographics significantly increases the perceived manipulateness of COVID-19 misinformation. These results are similar (and significant) in all three countries; see Table S6 for a full overview as well as item-level statistics. Figure 7 shows these results in a violin plot.

For real news, we also find a significant effect of condition on the pre–post difference in perceived manipulateness ( $F(2,1774) = 42.73$ ,  $p < 0.001$ ,  $\eta^2 = 0.046$ ). A Tukey post-hoc comparison shows that the perceived manipulateness of real news is significantly higher in the *Go Viral!* condition than both the control condition ( $M = 0.51$  vs  $M = 0.15$ ,  $M_{diff} = 0.36$ , 95% CI (0.25–0.46),  $p_{tukey} < 0.001$ ,  $d = 0.45$ ) and the Infographics condition ( $M = 0.51$  vs  $M = 0.14$ ,  $M_{diff} = 0.37$ , 95% CI [0.26, 0.48],  $p_{tukey} < 0.001$ ,  $d = 0.45$ ). However, the Infographics condition does not differ significantly from the control condition ( $M = 0.14$  vs  $M = 0.15$ ,  $M_{diff} = 0.01$ , 95% CI (–0.01 to 0.11),  $p_{tukey} = 0.950$ ,  $d = 0.02$ ). These results are similar across countries; see Table S6. We thus find partial support for hypothesis **H<sub>3</sub>**: playing *Go Viral!* initially increases

the perceived manipulateness of COVID-19 misinformation, but also of real news (although the effect size is descriptively smaller than for misinformation). The UNESCO infographics, on the other hand, only show this effect for misinformation (albeit to a lesser degree than *Go Viral!*), but not for real news.

To test hypothesis **H<sub>7a</sub>** and **H<sub>7b</sub>** for the manipulateness measure, we conducted a repeated measures one-way ANOVA with condition as the between-subjects factor and time (pre – post – follow-up) as the within-subjects factor, for the UK sample. Doing so shows a significant effect of time  $\times$  condition on the perceived manipulateness of misinformation ( $F(4,1206) = 5.85$ ,  $p < 0.001$ ,  $\eta^2 = 0.004$ ). Specifically, in the one-week follow-up, UK participants in the *Go Viral!* condition rated COVID-19 misinformation as significantly more manipulative than the control group ( $M = 5.96$  vs  $M = 5.64$ ,  $M_{diff} = 0.32$ , 95% CI (0.06–0.59),  $p_{tukey} = 0.011$ ,  $d = 0.30$ ). The Infographics condition, however, did not differ significantly from the control group in the follow-up ( $M = 5.73$  vs  $M = 5.64$ ,  $M_{diff} = 0.09$ , 95% CI (–0.15 to 0.32),  $p_{tukey} = 0.64$ ,  $d = 0.08$ ). As a (non-preregistered) robustness check, we conducted a between-subjects ANCOVA with pre-test manipulateness as the covariate and manipulateness in the follow-up as the dependent variable. Doing so shows a significant effect of condition on perceived manipulateness ( $F$



**Figure 8.** Bar graphs of perceived manipulateness of fake news and real news (UK only), by condition, for the pre-test (T1), post-test (T2) and 1-week follow-up (T3).

(2,602) = 5.54,  $p = 0.004$ ,  $\eta^2 = 0.013$ ), in that when controlling for the pre-test, participants in the *Go Viral!* condition find COVID-19 misinformation significantly more manipulative in the follow-up than the control group ( $M_{diff} = 0.29$ , 95% CI (0.07–0.51),  $p_{tukey} = 0.005$ ,  $d = 0.27$ ) and the Infographics condition ( $M_{diff} = 0.26$ , 95% CI (0.04–0.49),  $p_{tukey} = 0.015$ ,  $d = 0.25$ ).

For real news, while a repeated measures between-subjects ANOVA shows a significant effect of time  $\times$  condition on the perceived manipulateness of real news ( $F(4,1206) = 4.62$ ,  $p = 0.001$ ,  $\eta^2 = 0.003$ ), there was no significant difference between conditions for real news manipulateness in the one-week follow-up ( $F(2,603) = 2.04$ ,  $p = 0.131$ ,  $\eta^2 = 0.007$ ), indicating that participants across conditions rated real news as equally manipulative in the follow-up study. In addition, a (non-preregistered) between-subjects ANCOVA with pre-test real news manipulateness as the covariate and real news manipulateness in the follow-up as the dependent variable gives no significant difference between conditions ( $F(2,602) = 2.35$ ,  $p = 0.10$ ,  $\eta^2 = 0.005$ ). We thus find support for hypothesis  $H_{7a}$  and  $H_{7b}$ : one week after the initial UK intervention, *Go Viral!* players continued to rate COVID-19 misinformation as significantly more manipulative than the control group and people who read the UNESCO infographics, whereas the initial scepticism of real news

that was observed among *Go Viral!* players was no longer detectable one week after the intervention. Figure 8 shows these results in a bar graph. See Figure S1 for a plot that includes the data jitter.

**Confidence.** For the confidence measure, a between-subjects ANOVA on the pre–post difference in confidence scores for misinformation is significant ( $F(2,1774) = 29.47$ ,  $p < 0.001$ ,  $\eta^2 = 0.032$ ), in that participants in the *Go Viral!* condition are significantly more confident after the intervention in their assessment of misinformation than the control group ( $M = 0.34$  vs  $M = 0.05$ ,  $M_{diff} = 0.29$ , 95% CI (0.20–0.38),  $p_{tukey} < 0.001$ ,  $d = 0.44$ ) and the Infographics condition ( $M = 0.34$  vs  $M = 0.14$ ,  $M_{diff} = 0.20$ , 95% CI (0.11–0.29),  $p_{tukey} < 0.001$ ,  $d = 0.29$ ).<sup>15</sup> In addition, participants in the Infographics condition are also significantly more confident in their assessment of misinformation manipulateness than the control group ( $M = 0.14$  vs  $M = 0.05$ ,  $M_{diff} = 0.09$ , 95% CI (0.002–0.18),  $p_{tukey} = 0.043$ ,  $d = 0.15$ ). These results are similar (and significant) in all three countries, see Table S8.<sup>16</sup>

For real news, a between-subjects ANOVA shows no significant difference between conditions for the pre–post difference in confidence scores ( $F(2,1774) = 1.58$ ,  $p = 0.206$ ,  $\eta^2 = 0.002$ ). This finding is similar in

all three countries; see Table S8. We thus find support for hypothesis **H<sub>4</sub>**: participants in both the *Go Viral!* and Infographics conditions become significantly more confident in their ability to assess the manipulateness of misinformation about COVID-19 but show no change for real news. These results are again similar in all three countries (see Table S8).

To test hypothesis **H<sub>7a</sub>** and **H<sub>7b</sub>** for the confidence measure, we conducted a repeated measures one-way ANOVA with condition as the between-subjects factor and time (pre – post – follow-up) as the within-subjects factor, which shows a significant effect of time  $\times$  condition on confidence to assess the manipulateness of misinformation ( $F(4,1206) = 1.19$ ,  $p = 0.009$ ,  $\eta^2 = 0.003$ ). While a Tukey HSD post-hoc comparison gives no significant difference between conditions (all  $p_s > 0.12$ ), a (non-preregistered) between-subjects ANCOVA with pre-test confidence as the covariate and follow-up confidence as the dependent variable is significant ( $F(2,602) = 4.44$ ,  $p = 0.012$ ,  $\eta^2 = 0.010$ ), so that participants in the *Go Viral!* condition are significantly more confident than the control condition in their assessment of the manipulateness of misinformation one week after the intervention when controlling for the pre-test ( $M_{diff} = 0.25$  95% CI [0.05, 0.45],  $p_{tukey} = 0.011$ ,  $d = 0.23$ ). The results for the Infographics condition compared to the control group ( $p_{tukey} = 0.15$ ) and the *Go Viral!* condition compared to the Infographics condition ( $p_{tukey} = 0.45$ ) are not significant. For real news, a (non-preregistered) between-subjects ANCOVA with pre-intervention confidence as the covariate and follow-up confidence as the dependent variable is significant ( $F(2,602) = 3.49$ ,  $p = 0.031$ ,  $\eta^2 = 0.008$ ), so that Infographics participants are significantly more confident in their assessment of the manipulateness of real news than *Go Viral!* participants ( $M_{diff} = 0.24$ , 95% CI (0.004–0.47),  $p_{tukey} = 0.045$ ,  $d = 0.21$ ), whereas we find no difference between the *Go Viral!* and control condition ( $p_{tukey} = 0.85$ ) or the Infographics and the control condition ( $p_{tukey} = 0.09$ ). We thus find support for hypothesis **H<sub>7a</sub>**: one week after the intervention, people who played *Go Viral!* remained significantly more confident in their ability to assess the manipulateness of misinformation, but not real news. In partial support of **H<sub>7b</sub>**, participants who read the UNESCO infographics remained more confident in their ability to assess the manipulateness of real news, but not misinformation.

**Sharing.** For the sharing measure, a between-subjects ANOVA on the pre–post difference in willingness to share misinformation with others is significant ( $F(2,1774) = 4.00$ ,  $p = 0.019$ ,  $\eta^2 = 0.004$ ), in that participants in the *Go Viral!* condition are significantly less likely to indicate being willing to share misinformation

after the intervention than the control group ( $M = -0.18$  vs  $M = -0.07$ ,  $M_{diff} = 0.11$ , 95% CI (0.014–0.19),  $p_{tukey} = 0.019$ ,  $d = 0.15$ ).<sup>17</sup> However, we find no significant difference between the Infographics condition and the control group ( $p_{tukey} = 0.12$ ) nor the *Go Viral!* condition ( $p_{tukey} = 0.71$ ). These effects are directionally similar but not significant in each individual country (see Table S10).

For real news, we find no significant pre–post difference for the sharing measure between conditions ( $F(2,1774) = 0.28$ ,  $p = 0.75$ ,  $\eta^2 = 0.0003$ ), with similar results across countries (see Table S10). We thus find partial support for hypothesis **H<sub>5</sub>**: in the pooled sample, participants who played *Go Viral!* were significantly less willing than a control group to share misinformation about COVID-19 with people in their network. However, we find no such effects for infographics participants, and the effect is only visible in the higher-powered pooled sample (but not in country-level analyses).<sup>18</sup>

To test hypothesis **H<sub>7a</sub>** and **H<sub>7b</sub>** for the sharing measure, we conducted a repeated measures one-way ANOVA with condition as the between-subjects factor and time (pre – post – follow-up) as the within-subjects factor, which shows no significant effect of time  $\times$  condition on the self-reported willingness to share misinformation ( $F(4,1206) = 0.94$ ,  $p = 0.44$ ,  $\eta^2 = 0.001$ ). In addition, a (non-preregistered) between-subjects ANCOVA with pre-test willingness to share misinformation as the covariate and willingness to share misinformation in the follow-up as the dependent variable gives no significant difference between conditions ( $F(2,602) = 1.50$ ,  $p = 0.22$ ,  $\eta^2 = 0.003$ ). Similarly, for real news, a repeated measures one-way ANOVA shows no significant effect of time  $\times$  condition on the self-reported willingness to share real news ( $F(4,1206) = 0.99$ ,  $p = 0.412$ ,  $\eta^2 = 0.001$ ). We thus find no support for hypothesis **H<sub>7a</sub>** and support for **H<sub>7b</sub>** for the sharing measure; one week after the intervention, there is no longer a difference between conditions in terms of the self-reported willingness to share either real news or misinformation about COVID-19.

**Sharing the intervention with others.** To test hypothesis **H<sub>6</sub>**, we conducted an independent samples *t*-test (non-preregistered) with condition (*Go Viral!* or Infographics) as the independent variable and willingness to share the game or infographics on social media as the dependent variable. We find that *Go Viral!* participants are significantly more willing than Infographics participants to share the intervention with people in their network ( $M_{goviral} = 3.95$  vs  $M_{infographics} = 3.58$ ,  $M_{diff} = 0.37$ , 95% CI [0.15, 0.61],  $p = 0.001$ ,  $d = 0.19$ , 95% CI (0.08–0.31)), in support of hypothesis **H<sub>6</sub>**.



**Threat.** An ANOVA with traditional threat as the dependent variable and condition (*Go Viral!*, Infographics, control) and threat/post-test order as the independent variables revealed a non-significant effect for the overall model ( $F(5,1771)=1.84$ ,  $p=0.35$ ). In contrast, an ANOVA with motivational threat as the dependent variable and experimental conditions and threat/post-test order as the independent variables, shows that the overall model is marginally significant ( $F(5,1771)=2.19$ ,  $p=0.053$ ). This effect is primarily driven by the experimental condition ( $F(2, 1771)=4.06$ ,  $p=0.017$ , partial  $\eta^2=0.005$ ). Specifically, a Tukey HSD post-hoc comparison shows that participants in the *Go Viral!* condition indicated higher motivational threat than participants in the Infographics condition ( $M_{goviral}=5.59$  vs  $M_{infographics}=5.43$ ,  $M_{diff}=0.16$ ,  $p_{tukey}=0.028$ , 95% CI (0.013–0.31),  $d=0.15$ ) and the control condition ( $M_{goviral}=5.59$  vs  $M_{control}=5.44$ ,  $M_{diff}=0.15$ ,  $p_{tukey}=0.039$ , 95% CI (0.006–0.29),  $d=0.14$ ). There was no significant difference between the control and Infographics condition ( $p_{tukey}=0.98$ ). These results support our exploratory hypothesis **H<sub>8</sub>**.

## Discussion

We find that both prebunking interventions significantly increase the perceived manipulateness of misinformation about COVID-19, compared to a control group. This result is in line with Study 1 and remained valid in a randomised controlled setting and across three different languages. *Go Viral!* participants rated misinformation about COVID-19 as significantly more manipulative one week after the intervention, and were also significantly more confident in their judgments and experienced more motivational threat to defend their attitudes. With regards to real news, unlike Study 1, we find ambiguous results in Study 2: playing *Go Viral!* increases the perceived manipulateness of real news immediately after the intervention (similar to findings by Guess et al., 2020), whereas this effect is not observed for the infographics. However, this scepticism of real news among *Go Viral!* participants dissipates entirely after one week, unlike for misinformation, with real news being rated as equally manipulative across conditions in the follow-up.<sup>19</sup> We found no differences between conditions for real news for the confidence and sharing measures.

## General discussion

Across two large-sample studies using different research designs, we find strong support that both active and passive prebunking interventions increase people's ability to spot misinformation about

COVID-19 in social media content. Additionally, in line with previous studies (Basol et al., 2020; Saleh et al., 2021), we find that prebunking interventions increase people's confidence in their ability to spot misinformation. Crucially, this increase is in the right direction, so that people only became more confident in their ability when they correctly rated misinformation as manipulative. For *Go Viral!* players, these two effects remain significant for at least one week after gameplay, even when presented with previously unseen misinformation about COVID-19, indicating robust support for a high degree of retention of the inoculation effect (Maertens et al., 2020). These results also speak to the relative benefit of active versus passive inoculation especially in terms of delaying decay over time. We also note that, at least descriptively, the active intervention yielded larger effect sizes for manipulateness and confidence assessments than the passive intervention ( $d=0.56$  vs  $d=0.17$  for misinformation manipulateness, and  $d=0.44$  vs  $d=0.15$  for confidence). Finally, people were significantly more willing to share the *Go Viral!* game with others in their social media network than the infographics, which points towards a potential relative benefit of active versus passive prebunking interventions.

With respect to people's willingness to share social media content about COVID-19, we find that playing *Go Viral!* significantly reduces willingness to share misinformation about the virus (in line with Roozenbeek and van der Linden, 2020). However, this finding is not significant at the country level (although directionally similar). Furthermore, this effect was no longer significant after one week, and we find no difference in sharing willingness for the UNESCO infographics. As such, the #ThinkBeforeSharing infographics finding is inconsistent with recent research showing that getting people to pause and think can help reduce sharing of false news online (Fazio, 2020). It is possible that flooring effects are at play (i.e. participants had relatively low willingness to share both misinformation and real news even in the pre-test). Another possibility is that the sample sizes for the individual countries were not large enough to detect a significant effect; for example, a post-hoc power analysis for the sharing measure with  $d=0.15$ ,  $\alpha=0.05$  and  $n=710$  (which is what we obtained for the UK sample) returns an achieved power of 0.41. It is therefore possible that larger samples are needed to find consistent effects of misinformation interventions on sharing intentions.

This study also adds to the ongoing debate about the extent to which anti-misinformation interventions influence people's assessment of real news (Guess et al., 2020; Pennycook et al., 2020; Roozenbeek et al., 2020a). Our findings are somewhat ambiguous: while in Study 1 we find that playing *Go Viral!* does not

meaningfully affect people's assessment of real news, Study 2 suggests that *Go Viral!* players find real news about COVID-19 significantly more manipulative immediately after gameplay. Curiously, this effect is observed even for the items that were used in both studies. At the same time, confidence assessments and sharing intentions of real news are not affected by prebunking interventions (unlike those of misinformation), and any heightened scepticism of real news dissipates entirely one week after playing (unlike for misinformation). These results may be put into perspective with the decline of trust in news in recent years (Newman et al., 2020). Indeed, a recent cross-cultural study found that internet users' navigation on social media was based on a 'generalised scepticism' (Fletcher and Nielsen, 2018). Overall, our findings suggest that while prebunking interventions may (sometimes) influence people's assessment of real news (see also Guess et al., 2020), the presence and size of this effect varies substantially across studies and designs, and in the absence of established psychometric scales could be due to item effects rather than genuine underlying scepticism (Roozenbeek, et al., 2020a). We encourage further research on the implications of heightened scepticism of real news versus misinformation for truth discernment. Moreover, fact-checking is not without risk either (in some cases it can backfire, see Ecker et al., 2020; Krause et al., 2020), emphasising the question – as with medical treatments – whether the benefits of any (anti-misinformation) intervention outweigh potential side-effects.

Furthermore, inoculation theory has long regarded threat as integral in conferring attitudinal resistance, arguing that 'inoculation would be impossible without threat' (Compton and Pfau, 2005: 100–101). This observation is interesting because McGuire never explicitly measured threat himself, and a meta-analysis showed no significant relationship between threat and resistance, urging inoculation researchers to take a closer look at the role of threat (Banas and Rains, 2010). More recent scholarship suggests that 'motivational threat' – or threat in the form of motivation to defend oneself against persuasive attacks – is conceptually more consistent with inoculation than traditionally apprehensive threat (Banas and Richards, 2017; Richards and Banas, 2018). Our findings add to this debate by demonstrating promising effects of motivational threat (but not traditional threat) for the *Go Viral!* condition on the perceived manipulateness of misinformation. *Go Viral!* is therefore a promising step towards the development of interventions that motivate people to engage in attitudinal resistance without inadvertently heightening anxiety around the imminent attack or vulnerability of one's attitudes (Richards and Banas, 2018).

Like any study, ours has several limitations. First, our data is self-reported, and we were unable to assess how playing *Go Viral!* or reading the UNESCO infographics affects real-world behaviour. Second, we were only able to conduct a one-week follow-up in the UK. Third, it should be noted that only configural or weak invariance across the three language versions of this survey was established for the manipulateness, confidence, and sharing measures in Study 2 (both for the real and the misinformation items; see Tables S24–S28). However, the alpha levels for each construct were acceptable at the pooled and individual country level (0.72–0.92; see Table S5). Additionally, failing to reach a certain level of invariance should not prevent analyses so long as researchers note this limitation, which we do here (Putnick and Bornstein, 2016). Fourth, our study may not have achieved enough power to detect differences between conditions for the sharing measure, although judgments were made within a simulated social media setting (enhancing ecological validity). Fifth, in selecting the treatment comparisons, we favoured real-world generalisability over maximising internal validity for this study so future research may want to adopt a passive control that is more identical to the active condition across key parameters of interest.

## Conclusion

Across two large-sample studies, we provide strong cross-cultural evidence for the effectiveness of two short and easily scalable prebunking interventions to reduce susceptibility to misinformation about COVID-19. *Go Viral!*, a five-minute free-to-play browser game, positively impacts people's ability to identify misinformation about the virus for at least one week after playing, and significantly reduces intentions to share misinformation with others. We argue that prebunking constitutes a crucial step in the mitigation of misinformation about the pandemic. Finally, as the success of COVID-19 vaccination programmes worldwide depend in part on minimising the amount of unreliable information that surrounds them, our findings add to the emerging insight that interventions informed by behavioural science are a crucial tool to help mitigate the spread of misinformation.

## Acknowledgements

We wish to thank the Cabinet Office of the United Kingdom for funding the *Go Viral!* game and the World Health Organization and the United Nations for their support during and after the launch. We are also grateful to DROG and Gusmanson Design for their key role in the design, launch and hosting of the game, and to UNESCO (and specifically Isabel Tamoj) for translating the misinformation infographics to German.

## Author contributions

Melisa Basol and Jon Roozenbeek were primarily and equally responsible for the study design, data collection and analysis, visualisations and write-up, under the supervision of Sander van der Linden and with assistance from William P. McClanahan, Fatih Uenal and Manon Berriche. Manon Berriche and Fatih Uenal were responsible for the French and German translations of the *Go Viral!* game, respectively. William P. McClanahan was responsible for the invariance testing. All authors contributed to editing the manuscript.







## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

We are grateful for financial support from the University of Cambridge's ESRC IAA COVID-19 Rapid Response Fund and the United Kingdom's Cabinet Office.

## ORCID iDs

Melisa Basol  <https://orcid.org/0000-0003-1480-6462>  
 Jon Roozenbeek  <https://orcid.org/0000-0002-8150-9305>  
 Manon Berriche  <https://orcid.org/0000-0003-1381-8330>  
 Fatih Uenal  <https://orcid.org/0000-0002-8155-3066>  
 William P. McClanahan  <https://orcid.org/0000-0002-6604-3842>  
 Sander van der Linden  <https://orcid.org/0000-0002-0269-1744>

## Notes

1. See: [www.who.int/campaigns/connecting-the-world-to-combat-coronavirus/how-to-report-misinformation-online](http://www.who.int/campaigns/connecting-the-world-to-combat-coronavirus/how-to-report-misinformation-online)
2. Due to the low number of items, the averaged indices for the misinformation and real news items have relatively low internal consistency ( $\alpha_{\text{misinformation}} = 0.53$  and  $\alpha_{\text{realnews}} = 0.71$ ). We therefore also report item-level alongside averaged results.
3. See the Supplementary Analyses and Table S3 for the full Bayesian robustness testing.
4. The preregistrations can be found here: <https://aspredicted.org/28sr5.pdf> (UK), <https://aspredicted.org/fj5qh.pdf> (German), <https://aspredicted.org/d23y4.pdf> (French). Aside from the sample size (discussed in the 'Sample' section), the addition of hypothesis **H<sub>8</sub>**, and the addition of a non-preregistered ANCOVA analysis to test hypotheses **H<sub>7a</sub>** and **H<sub>7b</sub>** (as indicated in the 'Results' section), we report no significant deviations from the preregistration.
5. To stimulate a real social media environment as much as possible, we deliberately did not set a minimum time for participants to read through the infographics. Participants thus had to scroll through the page at their own pace, much like one may come across the infographics on one's Twitter feed or Facebook timeline.
6. The items can be found in the 'items' folder on the OSF: <https://osf.io/mbqwj/>
7. Infographics participants were asked whether the infographics contained a certain hashtag. *Go Viral!* participants were required to provide a completion code given at the end of the game and answer a question about the final scenario.
8. The presentation of threat items was counterbalanced to avoid order-effects on the post-rating task.
9. A confirmatory factor analysis was conducted for Richards and Banas' (2018) motivational threat measure. Following their use of Hu and Bentler's (1999) two-index criteria of comparative fit index ( $CFI > .95$ ) and the standardized root mean square residual ( $SRMR < .08$ ), the model demonstrated good fit,  $\chi^2(34) = 9.48$ ,  $p < .01$ ,  $CFI = .94$ ,  $SRMR = .048$ . Thus, the items exhibited unidimensionality.
10. Due to budgetary constraints, we were unable to do a follow-up for all three languages.
11. The full dataset for the pilot study, in which we pre-tested a set of 15 misinformation and 15 real items, can be found on the OSF: <https://osf.io/mbqwj/>. In total, 18 items (9 real and 9 misinformation) were used in the pilot study; 12 were rejected. For the follow-up study, we used six real and six misinformation items not tested in the pilot study.
12. For the German and French studies, we used 'German-speaking' and 'French-speaking' as inclusion criteria; pre-selecting for German or French nationality did not yield a large enough pool of Prolific participants.
13. We primarily report the results for the pooled sample (UK, French and German put together). The results per country are all directionally similar, but, where applicable, we report differences in significance levels. All item-level statistics for each outcome variable of interest (descriptive statistics and ANOVA results), pooled and by country, can be found in Tables S5 to S11. See Tables S24 to S28 for invariance testing between countries for both the real and misinformation items for the manipulateness, confidence, and sharing measures.
14. We report the eta squared (not partial eta squared) throughout the results section for study 2, except where specified otherwise. A reliability analysis gives acceptable internal consistency for the manipulateness measure for the misinformation items ( $M = 5.26$ ,  $SD = 0.94$ , Cronbach's  $\alpha = 0.77$ ) and the real items ( $M = 3.13$ ,  $SD = 1.02$ , Cronbach's  $\alpha = 0.81$ ), so we report the results for the average of both real and misinformation items. See also Table S7.
15. A reliability analysis shows good internal consistency for the confidence measure for the misinformation items ( $M = 5.34$ ,  $SD = 1.01$ , Cronbach's  $\alpha = 0.86$ ) and the real items ( $M = 4.89$ ,  $SD = 1.15$ , Cronbach's  $\alpha = 0.90$ ), so we report the results for the average of both misinformation items and real items. See also Table S9.
16. We also checked if participants became more confident in their assessment of misinformation about COVID-19 if they also *correctly* perceived misinformation to be more manipulative. To do so, we conducted an ANOVA with the pre-post difference in misinformation confidence as



the dependent variable, and condition and ‘manipulativeness-update’ (a binary variable that is positive if the pre-post manipulateness score for misinformation is positive and negative if this difference is negative or in the wrong direction) as fixed factors. Doing so shows a significant effect of condition  $\times$  manipulateness-update on misinformation confidence, in that participants across conditions who see misinformation as more manipulative after the intervention also show significantly greater confidence than participants who (incorrectly) see misinformation as *less* manipulative post-intervention. This effect is largest (descriptively) for *Go Viral!* participants; see Tables S29 and S30.

17. A reliability analysis shows good internal consistency for the sharing measure for both the misinformation items ( $M = 2.34$ ,  $SD = 1.31$ , Cronbach’s  $\alpha = 0.91$ ) and the real news items ( $M = 3.17$ ,  $SD = 1.25$ , Cronbach’s  $\alpha = 0.89$ ), so we report the results for the average of both real and misinformation items here. See also Table S11.
18. This may be the result of a floor effect given that average sharing intentions in online survey research are generally low (Pennycook et al., 2020).
19. We note that the real news items at follow-up were different items (to avoid item-memorisation effects).

## References

- Acerbi A (2019) Cognitive attraction and online Misinformation. *Palgrave Communications* 5(1): 1–7.
- Agley J, Xiao Y, Thompson EE, et al. (2020) COVID-19 misinformation prophylaxis: Protocol for a randomized trial of a brief informational Intervention. *JMIR Research Protocols* 9(12): e24383.
- Banas JA and Rains SA (2010) A meta-analysis of research on inoculation theory. *Communication Monographs* 77(3): 281–311.
- Banas JA and Richards AS (2017) Apprehension or motivation to defend attitudes? Exploring the underlying threat mechanism in inoculation-induced resistance to Persuasion. *Communication Monographs* 84(2): 164–178.
- Basol M, Roozenbeek J and 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): 2–9.
- BBC News (2020) Coronavirus: The fake health advice you should ignore. Available at: [www.bbc.co.uk/news/world-51735367](http://www.bbc.co.uk/news/world-51735367) (accessed 8 March 2021).
- Beriche M and Altay S (2020) Internet users engage more with phatic posts than with health misinformation on Facebook. *Palgrave Communications* 6(1): 1–9.
- Brady WJ, et al. (2017) Emotion shapes the diffusion of moralized content in social Networks. *Proc Natl Acad Sci USA* 114(28): 7313–7318.
- Compton J (2013) Inoculation theory. In: Dillard JP and Shen L (eds) *The SAGE Handbook of Persuasion: Developments in Theory and Practice*. 2nd edn. Thousand Oaks, CA: SAGE Publications, Inc., pp.220–236.
- Compton J (2019) Prophylactic versus therapeutic inoculation treatments for resistance to influence. *Communication Theory* 30(3): 330–343.
- Compton J and Pfau M (2005) Inoculation theory of resistance to influence at maturity: Recent progress in theory development and application and suggestions for future research. *Annals of the International Communication Association* 29(1): 97–145.
- Compton J and Pfau M (2009) Spreading inoculation: Inoculation, resistance to influence, and word-of-mouth communication. *Communication Theory* 19(1): 9–28.
- Compton J, van der Linden S, Cook J, et al. (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* e12602. DOI: 10.1111/spc3.12602
- Cook J, Lewandowsky S and Ecker UKH (2017) Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *Plos One* 12(5): 1–21.
- Dryhurst S, Schneider CR, Kerr J, et al. (2020) Risk perceptions of COVID-19 around the world. *Journal of Risk Research* 23(7–8): 994–1006.
- Eagly AH and Chaiken S (1993) *The Psychology of Attitudes*. Orlando, FL: Harcourt Brace Jovanovich.
- Ecker UKH, Lewandowsky S and Chadwick M (2020) Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire Effect. *Cognitive Research: Principles and Implications* 5(1): 41.
- Enders AM, Klofstad JE and Stoler C (2020) The different forms of COVID-19 misinformation and their consequences. *Harvard Kennedy School (HKS) Misinformation Review* 1(8): 1–21.
- Fazio L (2020) Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Misinformation Review* 1(2).
- Fazio L, Brashier NM, Payne BK, et al. (2015) Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General* 144(5): 993–1002.
- Fletcher R and Nielsen RK (2018) Generalised scepticism: How people navigate news on social media. *Information Communication and Society* 22(12): 1–19.
- Guess AM, Lerner M, Lyons B, et al. (2020) A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117(27): 15536–15545.
- Hu LT and Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6(1): 1–55.
- Imhoff R and Lamberty P (2020) A bioweapon or a hoax? The link between distinct conspiracy beliefs about the coronavirus disease (COVID-19) outbreak and pandemic Behavior. *Social Psychological and Personality Science* 11(8): 1110–1118.
- Ivanov B, Miller CH, Compton J, et al. (2012) Effects of postinoculation talk on resistance to influence. *Journal of Communication* 62(4): 701–718.
- Ivanov B, Sellnow T, Getchell M, et al. (2017) The potential for inoculation messages and postinoculation talk to minimize the social impact of politically motivated acts of



- violence. *Journal of Contingencies and Crisis Management* 26(4): 414–424.
- Jolley D and Douglas KM (2017) Prevention is better than cure: Addressing anti-vaccine conspiracy Theories. *Journal of Applied Social Psychology* 47(8): 459–469.
- Jolley D and Paterson JL (2020) Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British Journal of Social Psychology* 2020; 59: 628–640.
- Krause NM, Freiling I, Beets B, et al. (2020) Fact-checking as risk communication: The multi-layered risk of misinformation in times of COVID-19. *Journal of Risk Research* 1466–4461.
- Kucharski A (2020) Misinformation on the coronavirus might be the most contagious thing about it. *The Guardian*, 8 February. Available at: [www.theguardian.com/commentisfree/2020/feb/08/misinformation-coronavirus-contagious-infections](http://www.theguardian.com/commentisfree/2020/feb/08/misinformation-coronavirus-contagious-infections) (accessed 26 April 2021).
- Lewandowsky S and van der Linden S (2021) Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*. Epub ahead of print. DOI: 10.1080/10463283.2021.1876983. <https://www.tandfonline.com/doi/full/10.1080/10463283.2021.1876983>
- Loomba S, et al. (2021) Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour* 2021; 5(3): 337–348.
- Maertens R, Roozenbeek J, Basol M, et al. (2020) Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied* 27(1): 1–16.
- McGuire WJ (1964) Inducing resistance against persuasion: Some contemporary approaches. *Advances in Experimental Social Psychology* 1: 191–229.
- McGuire WJ and Papageorgis D (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: 326–332.
- Miller CH, Ivanov B, Sims J, et al. (2013) Boosting the potency of resistance: Combining the motivational forces of inoculation and psychological reactance. *Human Communication Research* 39(1): 127–155.
- Miller JM (2020) Do COVID-19 conspiracy theory beliefs form a monological belief system? *Canadian Journal of Political Science* 53: 319–326.
- Newman N, Fletcher R, Schulz A, et al. (2020) Reuters Institute digital news report 2020. Available at: [https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR\\_2020\\_FINAL.pdf](https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf) (accessed 9 December 2020).
- O’Keefe DJ (2015) Message generalizations that support evidence-based persuasive message design: Specifying the evidentiary requirements. *Health Communication* 30(2): 106–113.
- Peer E, Brandimarte L, Samat S, et al. (2017) Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70: 153–163.
- Pennycook G, McPhetres J, Zhang Y, et al. (2020) Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science* 31(7): 770–780.
- Putnick DL and Bornstein MH (2016) Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review* 41: 71–90.
- Richards AS and Banas JA (2018) The opposing mediational effects of apprehensive threat and motivational threat when inoculating against reactance to health promotion. *Southern Communication Journal* 83(4): 245–255.
- Roozenbeek J and van der Linden S (2018) The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research* 22(5): 570–580.
- Roozenbeek J and van der Linden S (2019) Fake news game confers psychological resistance against online misinformation. *Humanities and Social Sciences Communications* 5(65): 1–10.
- Roozenbeek J and van der Linden S (2020) Breaking harmony square: A game that “inoculates” against political misinformation. *The Harvard Kennedy School (HKS) Misinformation Review* 1(8). DOI: 10.37016/mr-2020-47.
- Roozenbeek J, Maertens R, McClanahan W, et al. (2020a) Differentiating item and testing effects in inoculation research on online misinformation. *Educational and Psychological Measurement* 81(2): 1–23.
- Roozenbeek J, Schneider CR, Dryhurst S, et al. (2020b) Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science* 7: 201199.
- Roozenbeek J, van der Linden S and Nygren T (2020) Prebunking interventions based on “inoculation” theory can reduce susceptibility to misinformation across cultures. *The Harvard Kennedy School (HKS) Misinformation Review* 1(2). DOI: 10.37016/mr-2020-008.
- Rucker DD and Petty RE (2004) When resistance is futile: Consequences of failed counterarguing for attitude certainty. *Journal of Personality and Social Psychology* 86(2): 219–235.
- Saleh N, Roozenbeek J, Makki FA, et al. (2021) Active inoculation boosts attitudinal resistance against extremist persuasion techniques – A novel approach towards the prevention of violent extremism. *Behavioural Public Policy*. Epub ahead of print 2021. DOI: 10.1017/bpp.2020.60. <https://www.cambridge.org/core/journals/behavioural-public-policy/article/active-inoculation-boosts-attitudinal-resistance-against-extremist-persuasion-techniques-a-novel-approach-towards-the-prevention-of-violent-extremism/EC1BF962A2B0012982BBB1507288E188>
- Schaeffer K (2020) Nearly three-in-ten Americans believe COVID-19 was made in a lab. *Pew Research Center*, 8 April. Available at: [www.pewresearch.org/fact-tank/2020/04/08/nearly-three-in-ten-americans-believe-covid-19-was-made-in-a-lab/](http://www.pewresearch.org/fact-tank/2020/04/08/nearly-three-in-ten-americans-believe-covid-19-was-made-in-a-lab/) (accessed 26 April 2021).
- Tormala ZL (2016) The role of certainty (and uncertainty) in attitudes and Persuasion. *Current Opinion in Psychology* 10: 6–11.

- Tormala ZL and Petty RE (2002) What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology* 83(6): 1298–1313.
- UNESCO (2020) #ThinkBeforeSharing – Stop the spread of conspiracy theories. Available at: <https://en.unesco.org/themes/gced/thinkbeforesharing> (accessed 27 November 2020).
- Van Bavel JJ, Baicker K, Boggio PS, et al. (2020) Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour* 4(5): 460–471.
- Van der Linden S and Roozenbeek J (2020) Psychological inoculation against fake news. In: Greifenader R, Jaffé ME, Newman EJ, et al. (eds) *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*. London, UK: Psychology Press.
- Van der Linden S, Leiserowitz A, Rosenthal S, et al. (2017b) Inoculating the public against misinformation about climate change. *Global Challenges (Hoboken, NJ)* 1(2): 1600008.
- Van der Linden S, Maibach E, Cook J, et al. (2017a) Inoculating against misinformation. *Science* 358(6367): 1141–1142.
- Van der Linden S, Roozenbeek J and Compton JA (2020) Inoculating against fake news about COVID. *Frontiers in Psychology* 11(566790): 19.
- Van der Linden S, Roozenbeek J, Maertens R, et al. (2021) How can psychological science help counter the spread of fake news? *Spanish Journal of Psychology* 24: e25.
- van Doorn J, van den Bergh D, Böhm U, et al. (2020) The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review* 1–14. DOI: 10.3758/s13423-020-01798-5
- Vraga EK and Bode L (2020) Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication* 37(1): 136–144.
- Wood MJ, Douglas KM and Sutton RM (2012) Dead and alive: Beliefs in contradictory conspiracy theories. *Social Psychological and Personality Science* 3(6): 767–773.
- World Health Organization (2020a) Coronavirus disease (COVID-19) advice for the public: Mythbusters. Available at: [www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters](http://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters) (accessed 5 May 2020).
- World Health Organization (2020b) *Countering misinformation about COVID-19*. Available at: [www.who.int/news-room/feature-stories/detail/countering-misinformation-about-covid-19](http://www.who.int/news-room/feature-stories/detail/countering-misinformation-about-covid-19) (accessed 29 June 2020).
- Zarocostas J (2020) How to fight an Infodemic. *The Lancet* 395(10225): 676.