Secondary Publication

Leder, Johannes; Schellinger, Lukas Valentin; Maertens, Rakoen; u. a.

Feedback exercises boost discernment of misinformation for gamified inoculation interventions

Date of secondary publication: 27.10.2025

Version of Record (Published Version), Article

Persistent identifier: urn:nbn:de:bvb:473-irb-110997x

Primary publication

Leder, Johannes; Schellinger, Lukas Valentin; Maertens, Rakoen; u. a. (2024): Feedback exercises boost discernment of misinformation for gamified inoculation interventions, in: Journal of experimental psychology / General, Washington, DC: American Psychological Association (APA), Vol. 153, Nr. 8, pp. 2068–2087, doi: 10.1037/xge0001603.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2024 The Author(s)
ISSN: 0096-3445

2024, Vol. 153, No. 8, 2068–2087
https://doi.org/10.1037/xge0001603

Feedback Exercises Boost Discernment of Misinformation for Gamified Inoculation Interventions

Johannes Leder[1], Lukas Valentin Schellinger[1], Rakoen Maertens[2], Sander van der Linden[3], Breanne Chryst[4], and Jon Roozenbeek[3]

[1] Department of Psychology, University of Bamberg
[2] Department of Experimental Psychology, University of Oxford
[3] Department of Psychology, University of Cambridge
[4] Cambridge Assessment, Cambridge, Cambridgeshire, United Kingdom

Gamification is a promising approach to reducing misinformation susceptibility. Previous research has found that "inoculation" games such as *Bad News* and *Harmony Square* help build cognitive resistance against misinformation. However, recent research has offered two important nuances: a potentially inadvertent impact of such games on people's evaluation of non-misinformation ("real news") and exponential decay over time if no memory-strengthening exercise is provided. We address these issues in two preregistered lab experiments (N_1 = 191, N_2 = 321) and four quasi-experimental in-game surveys implemented in *Harmony Square* (N_3 = 559) and *Bad News* (N_4 = 2,558, N_5 = 419, N_6 = 882). In Experiments 1 and 2, we test if providing different types of feedback after playing *Bad News* enhances discriminative ability of misinformation and real news 1 week postgameplay and find that doing so resulted in homogeneously better accuracy at identifying both misinformation and non-misinformation compared with a control condition, which played *Bad News* without feedback. In Experiments 3–6, we implemented two different types of feedback exercises in the *Harmony Square* and *Bad News* games and find that this significantly boosts discernment compared with playing the game without a feedback exercise, primarily by improving accuracy at detecting real news. We confirm these results using signal detection theory. We conclude that feedback exercises boost the effectiveness of gamified misinformation interventions, likely due to an improved learning environment.

Public Significance Statement

Across six separate experiments, we show that gamified inoculation interventions aimed at countering misinformation significantly benefit from a brief assessment of learning outcomes and providing players with feedback about their performance. Doing so (a) resulted in a stronger veracity discernment effect, including 1 week postintervention, and (b) increased the likelihood that all participants benefit equally from the intervention while eliminating and even reversing any undue skepticism of non-misinformation (or real news). Our findings have implications for misinformation intervention design, as there appear to be substantial benefits to optimizing the learning environment.

Keywords: misinformation, gamification, inoculation theory, fake news, feedback

Supplemental materials: https://doi.org/10.1037/xge0001603.supp

Sharda Umanath served as action editor.

Jon Roozenbeek https://orcid.org/0000-0002-8150-9305

Experiment 1 was financed with internal funds from the Chair of Personality Psychology and Assessment at the University of Bamberg (Germany). Johannes Leder and Lukas Valentin Schellinger are grateful to Astrid Schütz for her support. Experiment 2 was funded by the British Academy awarded to Jon Roozenbeek (PF21\210010) and IRIS coalition (U.K. government, SCH-00001-3391). Rakoen Maertens, Sander van der Linden, and Jon Roozenbeek are grateful for funding from JITSUVAX (EU Horizon 2020, 964728). Experiments 3–6 required no funding, but the *Bad News* and *Harmony Square* games were funded by the University of Cambridge (United Kingdom) and the Global Engagement Center (United States). The funding sources had no involvement in the study design, data collection, data analysis, interpretation of data, writing, or decision to submit. A version of this publication was previously posted as a preprint on PsyArXiv (https://osf.io/7k2mt/).

Open access funding is provided by the University of Cambridge: This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0; https://creativecommons.org/licenses/by/4.0). This license permits copying and redistributing the work in any medium or format, as well as adapting the material for any purpose, even commercially.

Johannes Leder played a lead role in conceptualization, formal analysis, and resources and an equal role in data curation, funding acquisition, investigation, methodology, software, validation, visualization, writing–original draft, and writing–review and editing. Lukas Valentin Schellinger played a supporting role in writing–original draft and writing–review and editing and an equal role in conceptualization, data curation, formal analysis, and methodology. Rakoen Maertens played a supporting role in conceptualization, data curation, formal analysis, visualization, writing–original draft, and writing–review and editing and an equal role in investigation, methodology, and project administration. Sander van der Linden played a lead role in supervision, a supporting role in

continued

Misinformation is a challenge for modern democracies, which rely on an informed citizenry (Kuklinski et al., 2000). The spread of misinformation may undermine public health (Swire-Thompson & Lazer, 2020; van der Linden, 2022), attitudes toward science (van der Linden et al., 2017), and the integrity of elections (Gunther et al., 2019; Lewandowsky et al., 2017). One approach to countering this problem, known as "prebunking," is to prevent misinformation from being ingrained into one's memory to begin with (Lewandowsky & van der Linden, 2021; Roozenbeek et al., 2023).

The most prominent prebunking approach is to confer psychological resistance against misinformation through psychological inoculation (Compton, 2013; Compton et al., 2021; Traberg et al., 2022). Inoculation theory (McGuire, 1961b) posits that, much like injecting a weakened dose of a virus to help the immune system recognize potential attackers and organize an effective immune response, people build up "cognitive antibodies" when confronted with a "weakened dose" of misinformation that is preemptively refuted in a controlled setting (Compton, 2013; McGuire, 1961a, 1964; McGuire & Papageorgis, 1962; van der Linden, 2024).

In recent years, researchers have used online games as a vehicle for delivering more "active" (McGuire & Papageorgis, 1962) inoculation interventions where players engage in experiential learning, such as *Cranky Uncle* (Cook et al., 2023), *Spot the Troll* (Lees et al., 2023), *Harmony Square* (Roozenbeek & van der Linden, 2020), and *Bad News* (Roozenbeek & van der Linden, 2019, 2022). In these types of games, players are inoculated not against individual examples of misinformation but against the manipulation techniques that often underlie a broader spectrum of misinformation, for example, conspiratorial thinking, emotional manipulation, or logical fallacies such as whataboutism, false dilemmas, or ad hominem attacks (see Roozenbeek, van der Linden, et al., 2022; Roozenbeek, Traberg, & van der Linden, 2022). Several studies, reviews, and a recent meta-analysis have found that such inoculation games are generally effective at reducing belief in misinformation (Iyengar et al., 2022; Lu et al., 2023; Traberg et al., 2022).

At this juncture, it is useful to conceptually distinguish inoculation from other forms of prebunking and competence-boosting interventions. Broadly speaking, the first and most obvious difference between prebunking and debunking concerns the timing of the intervention, such that any intervention prior to misinformation exposure can be considered a *pre*bunk (van der Linden, 2024), including simple forewarnings (Bertolotti & Catellani, 2023). However, the second difference relates to whether or not an actual competence or skill is being conferred, which is not the case with preemptive interventions such as some types of accuracy prompts (Pennycook et al., 2020) but is true of interventions that provide people with digital or media literacy tips (Guess et al., 2020; Lutzke et al., 2019; McGrew, 2020; Scheibenzuber et al., 2021).

What distinguishes *inoculation* from the aforementioned approaches besides the timing and level of detail is the specific format: Inoculations include (a) a forewarning that someone might be targeted with an attempt to manipulate their opinion (this is meant to elicit people's *motivation* to defend themselves from an impending attack) and (b) rather than providing people solely with facts or tips, the intervention preemptively exposes people to weakened doses of a falsehood or the techniques used to produce falsehoods along with ways on how to identify and refute them (this is meant to provide people with the actual *ability* to resist the manipulation attempt). In other words, an intervention can only be considered true inoculation if it contains weakened doses (examples) of the misinformation or manipulation attempt (McGuire, 1964; van der Linden, 2024). As mentioned, one novel way of doing this is through simulations in a gamified environment.

Several challenges remain for the efficacy of gamified inoculation interventions (and misinformation interventions more generally). Maertens et al. (2021) found that the inoculation effect fades over time without regular rehearsal. In two separate studies, Maertens et al. (2024) and Capewell et al. (2024) found that this decay may set in rapidly if no memory-strengthening exercise is provided (e.g., in the form of an item rating task administered postintervention). In other words, implementing a short task within the intervention that requires people to actively apply the lessons from the intervention may be a crucial component in its effectiveness.

Furthermore, there is evidence that participant responses obtained in some *Bad News* game studies were heterogeneous. For example, in Maertens et al.'s (2021) study, the standard deviation for the inoculation (treatment) group was twice as large postintervention compared with the control group. This suggests that the inoculation effect conferred by playing the game might be limited to a smaller group of participants than intended. Moreover, it has been proposed that inoculation games could inadvertently make players unduly skeptical of non-misinformation without corrective feedback; a reanalysis of several studies on gamified inoculation showed that some of these games may not improve "veracity discernment," a measure of people's ability to distinguish misinformation from non-misinformation (Modirrousta-Galian & Higham, 2023). This is important, as increased skepticism of *all* information (not just misinformation) may be a suboptimal outcome of many misinformation interventions (see Hameleers, 2023).

To address these shortcomings, this study assesses the effect of including short feedback exercises at the end of inoculation games in terms of boosting (a) the inoculation effect (i.e., veracity discernment) and (b) their longevity (although this latter outcome measure is preliminary). By feedback we here mean "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007, p. 81). Feedback is one central mechanism for learning the adaptive response to ambiguous stimuli (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Pan & Rickard, 2018). Research suggests that providing feedback about one's performance on a relevant task may help by reinforcing correct responses and correcting errors (Butler et al., 2007; Smith &

writing–original draft and writing–review and editing, and an equal role in conceptualization. Breanne Chryst played a supporting role in data curation and an equal role in formal analysis, software, and validation. Jon Roozenbeek played a lead role in project administration and supervision and an equal role in conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, validation, visualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Jon Roozenbeek, Department of Psychology, University of Cambridge, Downing Street, CB2 3EB Cambridge, United Kingdom. Email: jjr51@cam.ac.uk

Kimball, 2010) and boost memorization (Bird et al., 2015). A meta-analysis by Rowland (2014) found that the testing effect (i.e., when learners are tested on relevant information rather than merely revising content) is stronger when followed by feedback. The underlying mechanism appears to be that prediction errors (the deviation of the observed outcome from the expected outcome) are fundamental for episodic memory (Jang et al., 2019), declarative memory formation (Greve et al., 2017), and metacognition (Butler et al., 2008). In practice, this may mean that discernment and memory of the intervention is strengthened if feedback is provided after an intervention, which may translate to (among other things) increased longevity, that is, reduced effect decay over time (Capewell et al., 2024).

In gamified interventions, it is not practical to provide detailed individual feedback to each participant about everything that they learned. In the present study, we therefore focus on providing feedback on participants' performance on a relevant outcome measure, in this case an item task (e.g., evaluating a series of true and false/misleading/unreliable tweets or headlines). In a large number of studies (see Roozenbeek & van der Linden, 2024, Chapter 7 for a review), this item task performance is taken as a key measure of whether a misinformation intervention "works." Specifically, in the present study, we test if feedback in the form of (a) information on whether their ratings were correct or incorrect (e.g., correctly identifying misinformation as such) and (b) information on which manipulation technique (if any) is used in an item can boost task performance.

In terms of how feedback fits with the inoculation analogy, we follow Compton's (2013) suggestion to expand on the analogy to further theoretical development. Just as immune cells are excellent students insofar that T cells are exposed to a process of trial and error (i.e., feedback) to ensure that they are trained appropriately in discerning between molecules from our own bodies versus foreign invaders (Segel & Bar-Or, 1999), we posit that the process of generating psychological resistance will benefit from a similar training by bundling resistance with accuracy motivations to enhance discernment between manipulative and nonmanipulative information.

Finally, several recent studies have explored the role of self-efficacy, which relates to people's belief in their ability to act in the ways necessary to achieve one's goals (Hopp, 2022; Paciello et al., 2023; Rasmussen et al., 2022). However, although Rasmussen et al. (2022) found that some misinformation interventions can boost self-efficacy, no studies have been conducted that look at whether self-efficacy can act as a mediator in the effectiveness of misinformation interventions. Moreover, in the context of feedback, positive experiences can enhance self-efficacy, whereas negative experiences diminish it (Bandura, 1997). Chan and Lam (2010) found that formative feedback (i.e., feedback that allows for subsequent additional opportunities to practice) after failure on a test was significantly less detrimental to students' self-efficacy than summative feedback (given after the learning process has already completed). In the present study, we therefore also sought to explore whether providing feedback about people's performance during a misinformation intervention can affect self-efficacy compared with when no feedback is provided.

For this study, we conducted six separate experiments: two preregistered lab studies on Prolific where we tested the efficacy of various types of feedback exercises administered at the end of the *Bad News* inoculation game, both immediately postgameplay and 1 week after, and four (non-preregistered) field tests with pre- and postsurveys implemented in the *Bad News* and *Harmony Square* games; these last four studies were quasi-experimental as it was not possible to randomly assign participants to play the games either with or without feedback (we instead collected data across different time points). Across all experiments, we find that including a feedback exercise at the end of misinformation interventions significantly and substantially boosts veracity discernment, that is, people's ability to distinguish misinformation from non-misinformation (defined in all experiments as participants' average ratings of real news minus ratings of misinformation, as measured on a 1–7 scale ranging from 1 [*very unreliable*] to 7 [*very reliable*]; see Maertens et al., 2021). In addition, we find that feedback also improves performance 1 week after playing the game. This effect is robust for various types of feedback exercises and across item sets. See Figure 1 for an overview of the design of Experiments 1 and 2.

Transparency and Openness

We report how we determined our sample sizes for Experiments 1 and 2, all data exclusions (if any), all manipulations, and all measures in each study, following Journal Article Reporting Standards guidelines (Kazak, 2018); see Supplemental Material A. Experiments 1 and 2 were preregistered at https://aspredicted.org/SXC_VZ4 (Experiment 1) and at https://aspredicted.org/SDW_TG7 (Experiment 2). For Experiments 3–6, we copied the exact hypotheses and analysis plans from previous similar studies. Analyses were conducted in R Markdown (Experiments 1 and 2) and R (Version 4.2.3; Experiments 3–6). All analysis scripts as well as both the raw and cleaned data sets, along with any other supplementary information necessary to replicate our findings, can be found on our Open Science Framework (OSF) page (https://osf.io/cnr6t/).

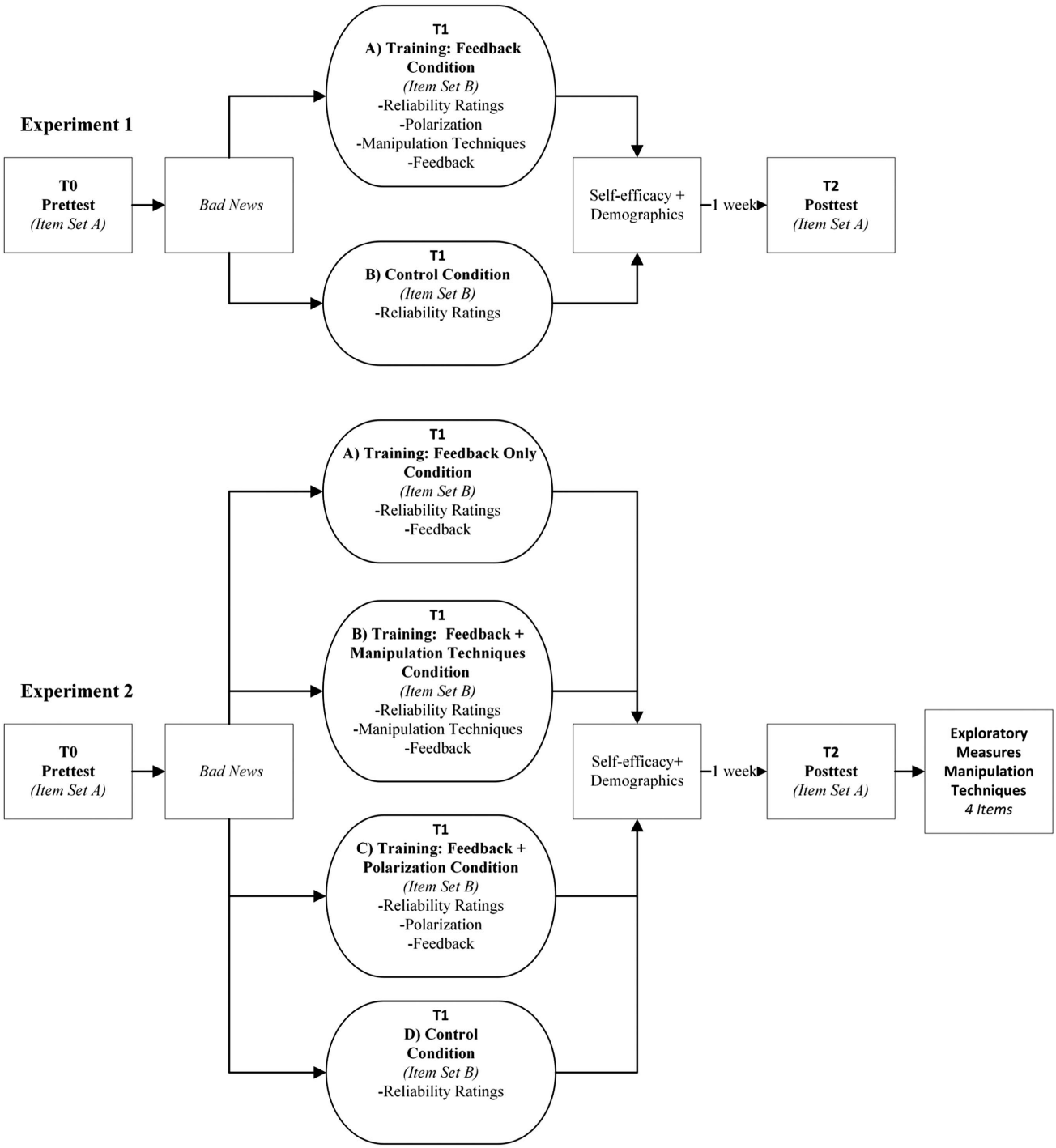
Experiment 1

In Experiment 1 (preregistration: https://aspredicted.org/SXC_VZ4), all participants completed an item rating task (consisting of a series of Twitter posts, some containing misinformation and some not) as a baseline measure (the pretest, T0), and then played *Bad News*. In the game, players roleplay as a "fake news tycoon" and set up their own news site and social media platform (this process is meant to reveal vulnerability to manipulation). Over the course of six levels, each representing a particular manipulation technique (impersonating fake accounts, leveraging negative emotions, polarizing audiences, spreading conspiracy theories, discrediting opponents, and trolling people for attention and outrage), players are tasked with gaining as large a following as possible by making use of weakened doses of each of these techniques in a controlled simulated setting. The game is interactive and choice-based, with players getting points (and followers) for making "correct" decisions (such as successfully spreading a fake news story and getting it picked up by the mainstream media). At the end of the game, players have built up a "fake news empire" and can challenge their friends. The *Harmony Square* game (see Experiment 3) is conceptually similar to *Bad News*; only in this game, players operate as a secretive disinformation operative, tasked with wreaking havoc in the peaceful community of Harmony Square by mounting an influence campaign during an election campaign.

After playing, all participants completed a different item rating task (of the same design but with new items, the posttest, T1). Specifically,

Figure 1

Flowchart Depicting the Procedure of Experiments 1 and 2



half of our participants (the feedback condition) first rated the reliability of a news headline (on a 1–7 scale) and then chose which of the six manipulation techniques from the *Bad News* game (e.g., conspiratorial reasoning) they believed was used in the headline. The feedback contained information about the manipulation technique and whether the headline was fake or real, along with whether the rating they gave to the headline was correct or incorrect. For the misinformation items, ratings below/above neutral (4) were rated as correct/incorrect, respectively. For the real news items, ratings above/below neutral were rated as correct/incorrect. We also included a feature where participants could ask for additional information about the specific manipulation techniques learned about in the game

(impersonation, emotional manipulation, polarization, conspiratorial thinking, discrediting/ad hominem attacks, and trolling). The aim of this feedback was to identify potential errors and show participants a strategy for how to better detect misleading content. The participants were guided to scrutinize each post for misleading cues and, based on this, to infer the reliability of news headlines. Seven days later (T2), participants were asked to again complete an item rating task, which contained the same items as the pretest (T0); performance at T2 is the central test for this study. Doing so also affords us a preliminary investigation into whether providing feedback can help reduce effect decay over time. For a detailed discussion on this topic, we refer to Capewell et al.'s (2024) and Maertens et al.'s (2024) studies. Finally, we explore whether self-efficacy plays a mediating role in the inoculation effect conferred by the game. We tested the following hypotheses:

- *Hypothesis 1a:* Playing *Bad News* and consecutive testing (without feedback) increase the likelihood to assess the reliability of misinformation items correctly after 1 week—the reliability decreases.
- *Hypothesis 1b:* Playing *Bad News* and consecutive testing (without feedback) increase the likelihood to assess the reliability of real news items correctly—the reliability increases.
- *Hypothesis 2a:* Playing *Bad News* with feedback immediately postgameplay increases the ability of participants to assess the reliability of misinformation items correctly—the reliability decreases.
- *Hypothesis 2b:* Playing *Bad News* with feedback immediately postgameplay increases the ability of participants to assess the reliability of real news items correctly—the reliability increases.
- *Hypothesis 3:* Self-efficacy mediates the effects of the treatment.

We decided to deviate from our original preregistration (https://aspredicted.org/SXC_VZ4), as we came to believe after data collection that the measurement at T1 was not necessarily indicative of an inoculation effect. Maertens et al. (2024) found that item rating tasks administered immediately after playing a gamified inoculation intervention serve as a memory-strengthening exercise in itself. In other words, the item task at T1 should be seen as part of the training, and we therefore cannot see the measurement at T1 as entirely separate from the intervention (*Bad News*). Because all participants rated the item set at T1, which contained different items than the item set administered at T0 and T2, we therefore tested all hypotheses comparing T0 and T2 (after 1 week), and the wording of Hypotheses 1 and 2 was adjusted accordingly. The preregistered regression tables (including results at T1) are reported in Supplemental Material F (Supplemental Table S10; see also Supplemental Figures S1–S3).

Method

Power Analysis and Sample

A G*Power analysis (Faul et al., 2007) was conducted with an effect size $f = 0.175$, $\alpha = .05$, and a power of 0.80 (for details see Supplemental Material B). Using these criteria, 196 participants were required to be able to detect a main effect. In the first wave, we recruited 200 participants via Prolific Academic. The final sample was smaller than expected after applying our preregistered exclusion criteria (see Supplemental Material A). The final sample consisted of 169 participants at T1 (i.e., immediately after completing *Bad News*). One week later at T2, a total of 152 participants completed the follow-up study (attrition rate: 11%). To reach the planned sample size of 200 participants, we conducted a second wave of data collection where 51 additional participants were recruited, with another 39 participants completing both parts of the study after applying our exclusion criteria. Doing so yielded a final sample size of $N = 191$. Of these, 49.2% describe themselves as female, 49.2% as male, and 1.6% as other. Participants' age consisted of four age groups: 65.4% 18–29 years, 31.4% 30–49 years, 2.6% over 50 years; and 0.5% under 18 years. The sample was international with participants from, for example, South Africa (40.3%), the U.K. (13.6%), and Mexico (5.2%); for sample descriptives see Supplemental Table S2. All demographic information item wordings can be found on the OSF page: Experiment_1_codeBook.pdf, pp. 7–9. A detailed description of all excluded participants and attention checks is provided in Supplemental Material A.[1]

Design and Procedure

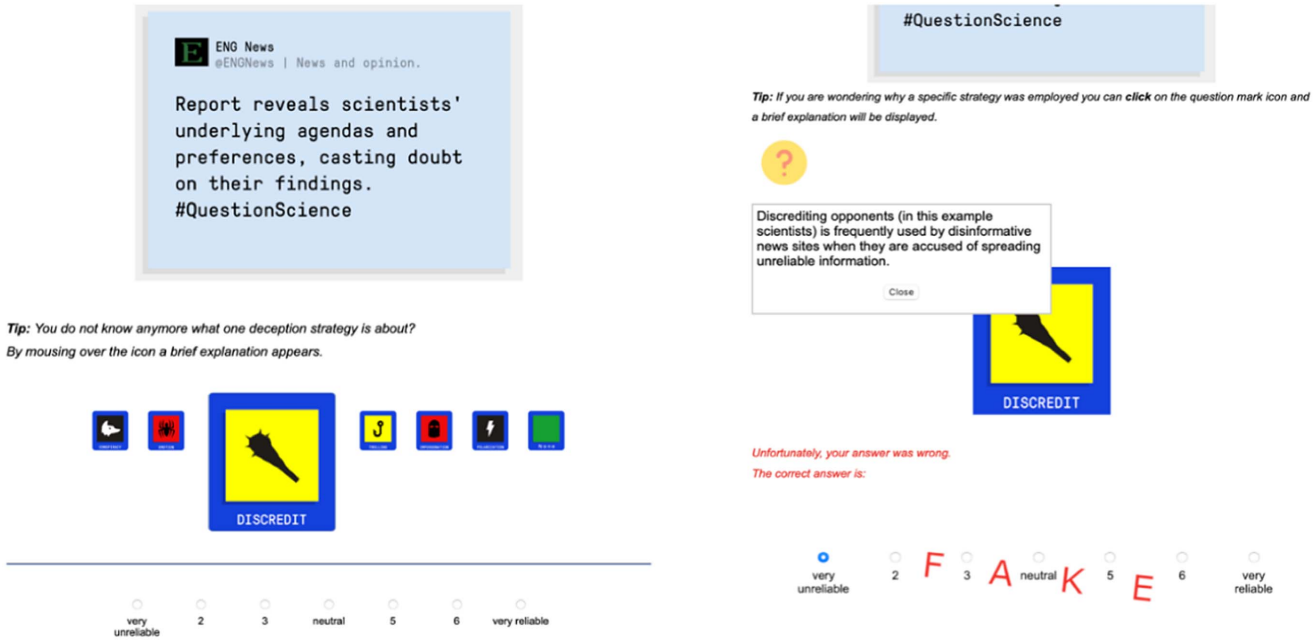
Participants were randomly assigned to either a control (*Bad News* only) or a feedback (*Bad News* + feedback) group. All participants were asked at three time points (T0, preintervention; T1, immediately postintervention; and T2, 7 days postintervention) to rate the reliability of a series of news items (in the form of Twitter posts). As feedback was provided on participants' reliability judgments in the feedback condition immediately postintervention (T1), we employed different item sets for the time points to control for potential memory effects: T0 = item set A; T1 = item set B; T2 = item set A (see Figure 1).

At T0, all participants rated the reliability of Twitter posts from item set A (12 items, nine misinformation; three real news; see the Measures section) and then played *Bad News* for approximately 15 min. The difference between the experimental group and the control group was that after completing the game (at T1), participants in the feedback group received feedback about their response after rating each item. In the feedback condition, immediately after playing the game, the items from item set B (nine misinformation; three real news) were presented, and participants were asked to rate each item's reliability and select which manipulation technique (from the game) was used in the item. They could also choose to get more information about each technique by clicking on a small symbol representing each technique (e.g., a spiked club representing the "discrediting opponents" technique; see Figure 2). Participants could then read additional information, taken from the Supplemental Material found on the *Bad News* website (DROG, 2018). The feedback included information about the manipulation technique and whether the news item contained misinformation or not. Positive feedback was provided only if participants gave extreme responses on the Likert scale, meaning they rated misinformation news items as very unreliable (1) or unreliable (2) and real news items as highly reliable

[1] Attention checks were taken from Shamon and Berning (2020) and adapted for the purpose of our study.

Figure 2

Screenshots Depicting the Layout of the Feedback Mechanism



Note. Left image shows the question page, followed by the actual feedback on the right. See the online article for the color version of this figure.

(7) or reliable (6). A label was shown over the Likert scale saying *real* or *fake*. For details on how much time participants spent on the feedback exercise at T1, see Supplemental Material H. One week postgameplay (at T2), participants again rated the items from item set A, this time without any feedback (see Figure 2).

Measures

Reliability Ratings. The items from item sets A and B were taken from previous research testing the efficacy of *Bad News* (Basol et al., 2020; Maertens et al., 2021; Roozenbeek & van der Linden, 2019). As in these studies, the reliability of the Twitter posts was measured on a Likert scale ranging from 1 (*very unreliable*) to 7 (*very reliable*), with 4 being *neutral*. One example of a misinformation item is, "the Bitcoin exchange rate is being manipulated by a small group of rich bankers. #InvestigateNow" (conspiracy technique). Item set A consisted of nine misinformation items and three real news items. For item set B, nine different misinformation items and three different real news items were used. Both sets cover all six manipulation techniques learned about in the *Bad News* game. At all time points, 12 items were shown. The scale properties are shown in Supplemental Table S8. The low internal consistency of the three real news items is in line with previous studies using the same item set (Maertens et al., 2021).

Self-Efficacy. We adapted a scale by Bong (2002) to assess English and mathematical self-efficacy. Players responded to two items on a 7-point Likert scale (1 = *not confident at all*, 7 = *very confident*; How confident are you that you can successfully identify fake news? How confident are you that you can successfully identify deception strategies frequently used in the production of fake news?). The items yielded good internal consistency ($\alpha = .79$) and were aggregated using the mean.

Exploratory Variables. We assessed a battery of personality characteristics,[2] but we do not analyze these in the present study due to space limitations.

Analytical Strategy and Deviations From the Preregistration

In Experiments 1 and 2, we used a statistical multiverse approach. This means we tested our preregistered hypotheses using different model families, including both Bayesian and frequentist approaches (Bürkner, 2018; Bürkner & Vuorre, 2019; Christensen, 2019). We ran a linear mixed model to account for repeated measurements. We included participants and items (nested in real news and misinformation) as random effects. The fixed effects were the feedback (control vs. feedback condition), time (pretest T0 vs. posttest T2) and item type (real news vs. misinformation), as well as their interactions. As preregistered, we report the models with and without controlling for age and gender. However, we did not preregister the model family that we would use for the main analyses nor if we would analyze the individual scores for each item or on aggregate. For the main analysis, we decided to use an ordinal regression approach estimating individual responses. Because the responses for the reliability ratings were on an ordinal scale, we report the Bayesian results here, as only Bayesian models do not require homogenous variances and using metric models for ordinal data can result in biased inferences (Liddell & Kruschke, 2018). We report the full results in Supplemental Table S10. See

[2] Interpersonal Trust (three items; Beierlein et al., 2012); Personal Need for Structure (10 items; Neuberg & Newsom, 1993); Preferences for Intuition and Deliberation (19 items; Betsch, 2004).

Supplemental Material B for further details about our analysis plan (e.g., about the regression equations used to test for main effects).

We deviated from our preregistration in our tests of Hypotheses 1 and 2, in that we stated that we would compare the interaction with time, with time having three levels (T0, T1, T2). We decided not to include the results for T1 in our hypothesis testing, as the key manipulation of interest (i.e., the feedback) occurred during the item rating task at T1, which may interfere with item rating task performance at T1. For this reason, we only included ratings at baseline (T0) and 7 days later (T2). See Supplemental Table S10 and Figures S1–S3 for the results for T1. Finally, for the Bayesian analyses, we did not preregister the priors. However, to regularize our estimates and allow for faster model convergence, we used weakly informative priors with $\mu = 0$ and $\sigma = 5$ for all fixed effects and intercepts.

Results

We report the results for the best-fitting ordinal Bayesian model with heterogeneous variance between groups and time (see Supplemental Table S10). We present estimates in the unit of standard deviation (similar to Cohen's *d*, which represents the difference between two means standardized to the unit of the standard deviation) as point estimates of odds ratios with 95% credible intervals (CIs).[3] See Supplemental Table S7 for a table with the mean reliability ratings and their 95% CIs at each time point. See Supplemental Material E (Supplemental Figures S1–S3) for item-level plots and for plots of reliability ratings aggregated and not aggregated to mean scale values.

Bad News Effect (Without Feedback)

Testing Hypothesis 1a, in the control group (*Bad News* without feedback), the participants rated misinformation as less reliable at T2 than at T0, $b = -0.14$, 95% CI [−0.27, −0.01], in support of Hypothesis 1a and confirming that *Bad News* increases the likelihood of participants rating misinformation items correctly 1 week after playing (although the effect is weak, possibly due to insufficient power). For Hypothesis 1b, control group participants rated real news items at T0 as more reliable than misinformation, $b = 1.65$, 95% CI [1.12, 2.18], and did not rate real news as significantly less or more reliable at T2, $b = .03$, 95% CI [−0.16, 0.23] (though we note the low internal consistency of the three real news items; see Supplemental Table S8). We thus fail to find support for Hypothesis 1b, as we predicted that the reliability of real news would *increase* postgameplay; instead, the participants rated real news as equally reliable before playing *Bad News* and 7 days later.

Bad News + Feedback Compared With Bad News Only

Testing Hypothesis 2a, in the feedback condition, the participants rated misinformation as less reliable after playing *Bad News* than the control group at T2 compared with T0, $b = -0.56$, 95% CI [−0.77, −.35]. This confirms that providing feedback postgameplay is more effective in conferring resistance against misinformation compared with only playing *Bad News*, supporting Hypothesis 2a. For Hypothesis 2b, real news was rated as *more* reliable in the feedback condition than the control group at T2 compared with T0, $b = 1.03$, 95% CI [0.69, 1.38]. These results support Hypothesis 2b. Furthermore, the variance of the feedback group was lower than in the control group at T2, $b = -0.40$, 95% CI [−0.51, −0.28], as indicated by the model parameters (for all coefficients see Supplemental Table S10). We plotted the marginal effects and the observed and predicted probabilities for each response category in reliability ratings in Figure 3.

Self-Efficacy Effect. Finally, as preregistered, we also tested if the difference between experimental conditions could be explained by the difference in self-efficacy (mediation Hypothesis 3). We did not observe a difference in self-efficacy between experimental conditions (see Figures S7 and S8 in Supplemental G). Including self-efficacy in the regression predicting responses at T2 did not change the estimates, suggesting that no variance between conditions in ratings is explained by self-efficacy (see Supplemental Table S12).

Discussion

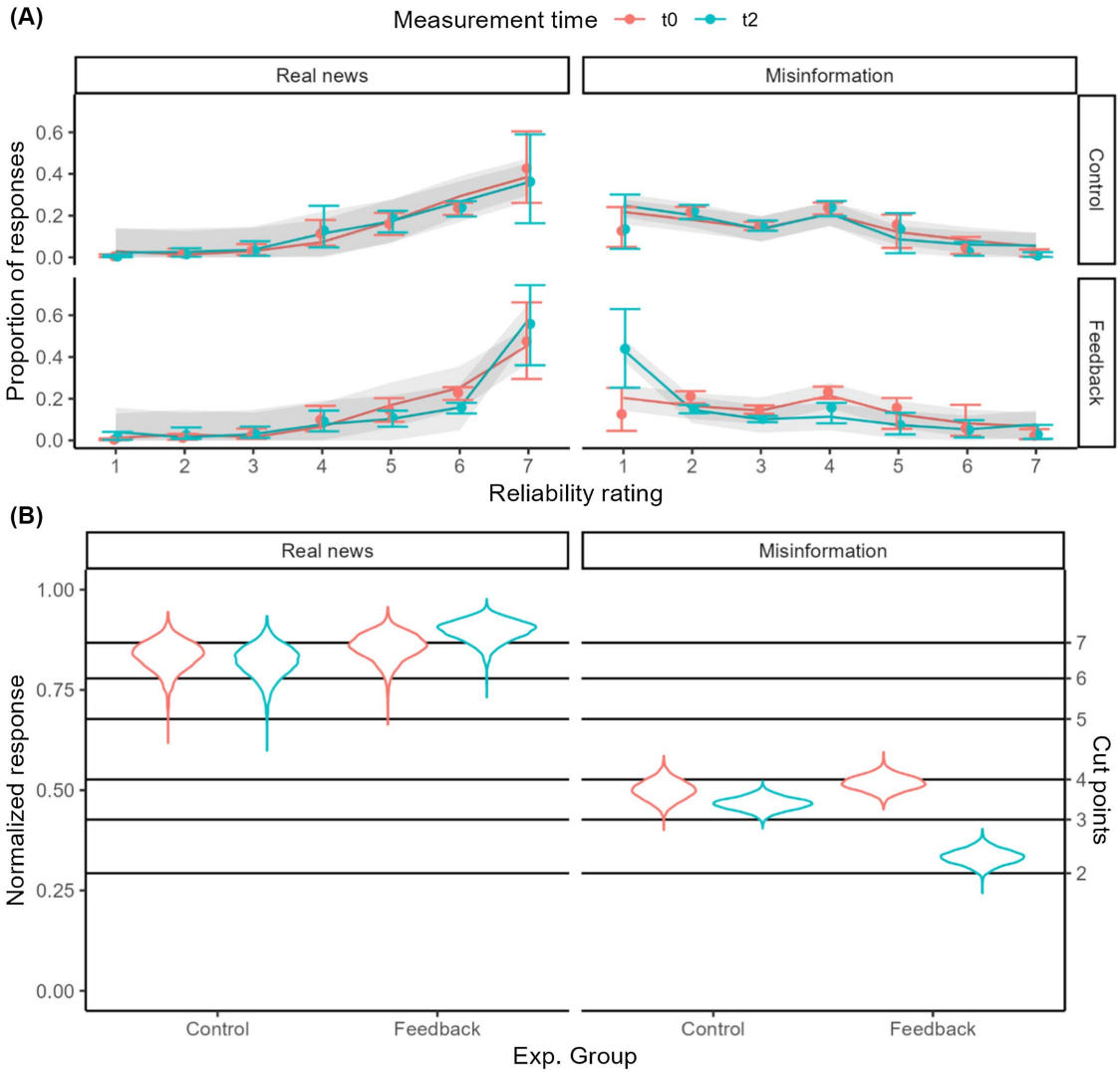
We tested if providing performance feedback after playing *Bad News* increases players' ability to correctly judge Twitter posts as real news or misinformation. In line with our hypotheses, 7 days after playing, participants in the feedback group increased their ability to rate misinformation as more unreliable and increased their trust in real news compared with participants who only played *Bad News* without feedback. Overall, the results suggest that the feedback had a positive influence on the efficacy of the intervention: A low reliability rating of misinformation was twice as likely 7 days postintervention if participants were given feedback about their performance. Moreover, in line with other research (Capewell et al., 2024; Maertens et al., 2024), our findings suggest that providing feedback can strengthen the effect of the intervention over time compared with when no feedback is provided. Finally, perceived self-efficacy to identify Twitter posts as real news or misinformation did not differ between groups, and it did not explain any variance in the reliability ratings, indicating that self-efficacy may play a limited role as a moderator of misinformation intervention effectiveness.

However, Experiment 1 had several limitations. Most importantly, the feedback condition contained several components that might drive the observed effects. First, the feedback informed players about the type of news (whether it contained misinformation or not). Second, the feedback showed the exemplary correct responses only as end-of-axis responses (1 or 7) visually, and positive feedback was provided only for extreme responses, (1) or (2) for misinformation and (6) or (7) for real news items, which could have resulted in more extreme response patterns. Furthermore, feedback entailed not only information about the ideal reliability rating but also about the manipulation technique used in each item. Finally, participants in the feedback condition were able to review the manipulation strategies at their own pace when making reliability judgments during training (immediately postgameplay), which could help to boost the effect further. We address these limitations in Experiment 2.

[3] Credible intervals (CIs), which are Bayesian confidence intervals (the numbers refer to the 2.5th and 97.5th percentiles of the posterior distributions), can be numerically similar to their frequentist counterparts (i.e., confidence intervals); however, they have an intuitive probabilistic interpretation, unlike confidence intervals, which are often mistakenly interpreted as probabilities (Hoekstra et al., 2014; Morey et al., 2016).

Figure 3

Experiment 1: Observed and Predicted Responses and Estimated Effects or Experimental Group, Time Point (t0 = Pregameplay, t2 = 1 Week Postgameplay), and Item Type (Real News or Misinformation)



Note. Plot A shows the reliability ratings observed with the line and ribbon (95% CI). The point and error bars show the model prediction based on a 95% highest density interval. Plot B violins show the posterior distribution of the beta coefficients in relation to the normalized responses and the cutoff points, based on 4,000 draws. Control = *Bad News* without feedback; CI = confidence interval; Exp. Group = experimental group. See the online article for the color version of this figure.

Experiment 2

In Experiment 2 (preregistration: https://aspredicted.org/SDW_TG7), we administered three separate treatment conditions: feedback only, feedback with manipulation technique choice, and feedback with polarization. In the feedback-only condition, participants were only told if their rating was correct or incorrect. Ratings below/above neutral (4) were rated as correct/incorrect, respectively. In the feedback + strategy condition, the participants also had to identify the manipulation technique used in each item and were able to read the information about the manipulation techniques again if they chose to. In the feedback + polarization condition, feedback included a correct extreme response: low (1) for misinformation and high (7) for real news (so rather than only indicating if the participant's response was correct or incorrect, participants in this condition were also shown the "correct" [extreme] rating for each item, 1/7 for misinformation items and 7/7 for real news). We preregistered the following hypotheses:

- *Hypothesis 1a:* Playing *Bad News* increases the likelihood to assess the reliability of misinformation items correctly—the reliability decreases.
- *Hypothesis 1b:* Playing *Bad News* increases the likelihood to assess the reliability of real news items correctly—the reliability increases.

- *Hypothesis 2a:* Playing *Bad News* with feedback immediately postgameplay increases the ability of participants to assess the reliability of misinformation news items correctly—the reliability decreases.
- *Hypothesis 2b:* Playing *Bad News* with feedback immediately postgameplay increases the ability of participants to assess the reliability of real news items correctly—the reliability increases.
- *Hypothesis 3a:* The effect of simple feedback (feedback only) is weaker than the effect of feedback + polarization.
- *Hypothesis 3b:* The effect of simple feedback (feedback only) is weaker than the effect of feedback + manipulation technique information.

Method

A power analysis was conducted to test Hypothesis 2. As the responses were on a Likert scale, we used an ordinal regression approach (see Liddell & Kruschke, 2018). We simulated results for the estimated regression coefficients from Experiment 1 with N = 1,000 iterations and calculated the proportion of $p < .05$. We assumed an effect size of $\beta = 1.89$ (logit scale) for the feedback effect at T1.[4] We assumed that the effect would be 50% of the effect size from Experiment 1, so we simulated the interaction of Time × Real news × Feedback with $\beta = .90$ and $\alpha = .05$. To achieve power within a 95% CI of $1 - \beta = [.96, .98]$ (meaning power is between 96% and 98%), our simulation suggested that 140 participants (n = 70 per condition) would be required. Details for the simulation can be found in Supplemental Material B. The script and resulting power curves from the power analysis are found in our OSF (https://osf.io/cnr6t/) repository ("Power Simulation" folder). As Experiment 2 has four experimental conditions, we therefore aimed to collect data from 280 participants (i.e., n = 70 per group) from Prolific Academic. Taking into account a potential attrition of 20%, we sampled 350 participants at T0. The sample consisted of U.K. residents only. As preregistered (see Supplemental Material A), once 350 participants had successfully completed the first part of the study, we stopped the data collection. In total, 312 participants completed Part 1 and Part 2, that is, a slight oversampling. Of these 65.7% were female, 33.7% were male, and 0.6% categorized themselves as "other." The average age was 39.9 years (SD = 12.4).[5] For sample descriptives see Supplemental Material C and Supplemental Table S3. Other than the experimental conditions (see above), the study design, measures, deviations from the preregistration, and exclusion criteria were the same as in Experiment 1.[6] The preregistered analysis plan was the same as the approach used in Experiment 1 as well; see Supplemental Material B for more details.

Exploratory Variables

We included one exploratory measure not included in Experiment 1, which we will discuss here: After all participants completed T2, they were again shown four misinformation items. For each item, participants were asked to identify which manipulation technique was used. The misinformation items were taken from item set A, and the Likert scale was replaced by categorical response options: impersonation, emotion, polarization, conspiracy, trolling, discrediting, or none. We did not find a significant effect in the identification of the correct techniques between conditions (feedback only, feedback + manipulation technique information, feedback + polarization, control); see Supplemental Material H for more details. However, when comparing performance across all conditions with an additional pretest, feedback type showed the strongest effect ($\beta = 1.47$, 95% CI [1.14, 1.90], $p = .003$), and participants in the feedback + manipulation technique information group had the highest percentage of correct responses.

Results

As in Experiment 1, we report the results of the best-fitting ordinal Bayesian model with heterogenous variance between groups and time (T0 and T2); see Supplemental Table S11. We again report the estimates in the unit of the standard deviation and report these as point estimates with 95% CIs. For the observed and predicted ratings as well as the distribution of the group means, we used the posterior distribution of the inverse logit transformed values. See Supplemental Table S7 for a table with the mean reliability ratings and their 95% CIs at each time point and Supplemental Table S8 for scale properties. See Supplemental Material E (Supplemental Figures S4–S6) for item-level plots at T0, T1, and T2 and for plots of reliability ratings, both aggregated and not aggregated to mean scale values.

Bad News Effect (Without Feedback)

Testing Hypothesis 1a, we observe a decrease in reliability ratings for misinformation in the control (*Bad News* without feedback) group when comparing T0–T2, $b = -0.35$, 95% CI [−0.50, −0.19]. These results support Hypothesis 1a. Real news items received high reliability ratings compared with misinformation at T0, $b = 2.17$, 95% CI [1.27, 2.96]. Testing Hypothesis 1b, control group participants (*Bad News* without feedback) rated real news as descriptively *more* reliable at T2, but the effect was nonsignificant as the 95% CI includes zero, $b = 0.22$, 95% CI [−0.01, 0.45]; these results do not support Hypothesis 1b, that is, playing *Bad News* does not increase the likelihood of participants identifying real news items correctly postgameplay.

Bad News + Feedback Compared With Bad News Only

Testing Hypothesis 2a, we found that participants in the feedback-only condition ($b = -0.35$, 95% CI [−0.59, −0.13]) and the feedback + polarization condition ($b = -0.71$, 95% CI = [−0.94, −0.45]) rated misinformation as less reliable than the control group (*Bad News* without feedback). However, for the feedback + manipulation technique condition, this effect only had a 90% certainty, and the mode of the 95% CI of the highest density interval includes zero, $b = -0.21$, 95% CI [−0.42, 0.03]. Overall, these

[4] As in Experiment 1, we no longer look at the inoculation effect at T1 (only at T2).

[5] In contrast to Experiment 1, an open textbox was used to measure age on a scale (a detailed overview of the demographic information survey is provided on the OSF repository, Experiment_2_CodeBook.pdf, pp. 7–9).

[6] With one exception: the exclusion criterion "participant has played *Bad News* before" was replaced with "participant entered the incorrect *Bad News* completion code (if in *Bad News* condition)."

results partly support Hypothesis 2a: The feedback-only and feedback + polarization conditions both boost the ability of participants to correctly identify misinformation, but this was not the case for the feedback + manipulation technique condition. For the real news items (Hypothesis 2b), we find that the increase in the perceived reliability of real news is larger in all three feedback conditions than the control group (so that all three feedback conditions rated real news as more reliable than the control group): feedback-only (b = 0.47, 95% CI [0.09, 0.79]), feedback + polarization (b = 0.74, 95% CI [0.34, 1.11], and feedback + manipulation technique information (b = 0.41, 95% CI = [0.09, 0.76]. These results support Hypothesis 1b.

The parameter estimates for the variances show that variance at T2 was lower than at T0, b = −0.09, 95% CI [−0.18, −0.01]; however, the variance was lower in all feedback conditions, but the effect was small, and all 95% CIs included zero. Taken together, all feedback conditions showed a stronger decrease in reliability ratings of misinformation and a stronger increase in the reliability of real news at T2 compared with T0. Overall, the results show that providing feedback about individuals' performance resulted in more homogeneous learning across participants, indicated by the fact that almost 50% of all respondents in those two groups chose the lowest reliability rating on the posttest (see Figure 4).

Feedback Effect Comparison. Testing Hypothesis 3, we calculated the difference between the posteriors for each draw and the resulting coefficient for each group reflecting the mean. We find that the feedback-only condition decreased the perceived reliability of misinformation less than the feedback + polarization condition, but the effect is not significant, b = 0.18, 95% CI [−0.05, 0.50]. However, both the feedback-only (b = −0.31, 95% CI [−0.57, −0.05]) and feedback + polarization (b = −0.50, 95% CI [−0.79, −0.25]) conditions resulted in significantly lower reliability ratings for misinformation than the feedback + manipulation technique condition. These results thus do not support Hypothesis 3a, as we hypothesized that simple feedback (feedback only) would result in a weaker effect than feedback + polarization, or Hypothesis 3b, as the effect of simple feedback is weaker than the effect of feedback + manipulation technique information.

Self-Efficacy Effect. Finally, as in Experiment 1, we did not observe a difference in self-efficacy between experimental conditions (see Figures S9 and S10 in Supplemental G). Including self-efficacy in the regression predicting responses at T2 did not change the estimates (see Supplemental Tables S13).

Discussion

In Experiment 2, we sought to isolate the effect of various types of feedback on participants' performance on item rating tasks 1 week after playing *Bad News*: only stating whether the participant's response to a (misinformation or real news) item was correct (feedback only), asking for additional information about which manipulation technique was used in the item (feedback + manipulation technique information), and providing a (correct) extreme response on the Likert scale (i.e., 1/7 for misinformation items and 7/7 for real news items) alongside the correct answer (feedback + polarization). Overall, we find that feedback exercises resulted in more homogeneous learning across experimental conditions compared with the control group, which only played *Bad News* without feedback. The feedback + manipulation technique information exercise was only successful at boosting the correct identification of real news and not misinformation. On the contrary, the feedback + polarization condition appeared to be descriptively the most successful, highlighted by the fact that almost 50% of participants in this condition rated all misinformation items as 1/7 on the reliability scale. However, this type of feedback might be impractical in the real world, as the goal of inoculation games is not to optimize responses on Likert scale rating scales but rather to boost a particular skill or competence such as correctly identifying manipulative content. In addition, this type of feedback exercise rewards black-and-white thinking (encouraging participants to rate items as either entirely unreliable or entirely reliable), which has little practical value, as real-world misinformation is rarely so unambiguous. We therefore test the effectiveness of the feedback-only and feedback + manipulation technique information feedback types in a "live" environment in Experiments 3–6.

Experiments 3–6

The goal of Experiments 3–6 was to test if two different feedback exercises (adapted versions of the feedback-only and the feedback + manipulation technique information exercises, respectively) boost performance on item rating tasks in the *Harmony Square* (https://www.harmonysquare.game/) and *Bad News* (https://www.getbadnews.com/) inoculation games. If doing so can be shown to substantially boost performance, this would indicate improved learning and alleviate concerns about response bias and potentially limited longevity (Capewell et al., 2024; Maertens et al., 2024; Modirrousta-Galian & Higham, 2023). Our (non-preregistered) hypotheses were either the same as for Experiments 1 and 2 or taken from previous studies with these games that used similar designs (Basol et al., 2021; Roozenbeek, Traberg, & van der Linden, 2022) and were as follows:

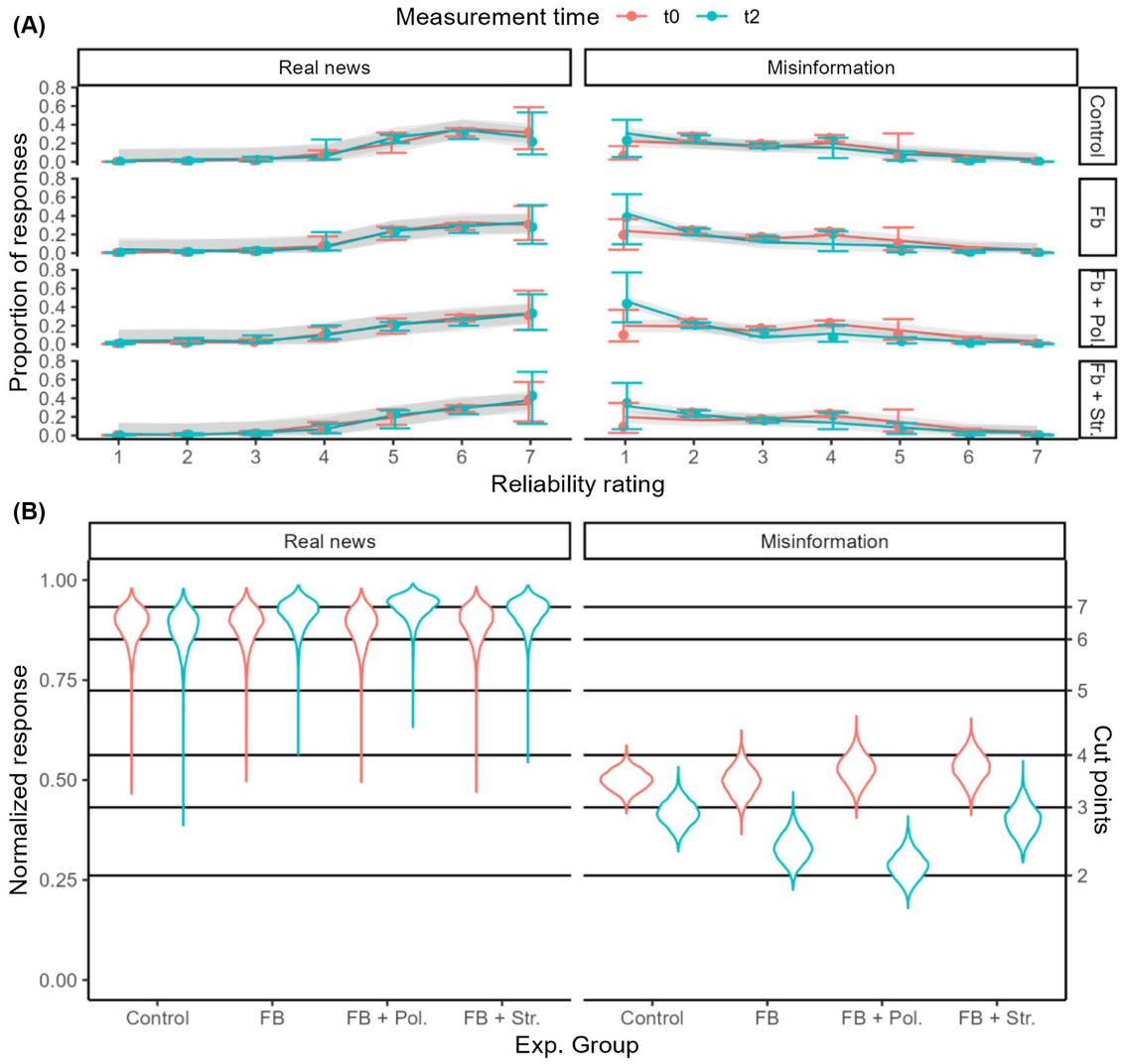
- *Hypothesis 1a:* Playing an inoculation game (*Harmony Square* or *Bad News*) without feedback reduces the perceived reliability of misinformation.
- *Hypothesis 1b:* Playing an inoculation game (*Harmony Square* or *Bad News*) without feedback does not reduce the perceived reliability of non-misinformation (real news).
- *Hypothesis 1c:* Playing an inoculation game (*Harmony Square* or *Bad News*) without feedback improves discernment between misinformation and real news.
- *Hypothesis 2a:* Feedback increases the ability of participants to assess the reliability of misinformation items correctly—the reliability decreases.
- *Hypothesis 2b:* Feedback increases the ability of participants to assess the reliability of real news items correctly—the reliability increases.
- *Hypothesis 2c:* Feedback increases discernment between misinformation and real news, compared with no feedback.

Method

The methodologies in Experiments 3 (*Harmony Square*) and 4–6 (all *Bad News*) are broadly similar: We implemented an item rating

Figure 4

Experiment 2: Observed and Predicted Responses and Estimated Effects or Experimental Group, Time Point (t0 = Pregameplay, t2 = 1 Week Postgameplay), and Item Type (Real News or Misinformation)



Note. Plot A shows the reliability ratings observed with the line and ribbon (95% CI). The point and error bars show the model prediction based on a 95% highest density interval. Plot B violins show the distribution of the posterior distribution of the betas in relation to the normalized responses and the cutoff points based on 4,000 draws. Control = *Bad News* without feedback; FB = feedback only; FB + Pol. = feedback + polarization; FB + Str. = feedback + manipulation technique information; CI = confidence interval; Exp. Group = experimental group. See the online article for the color version of this figure.

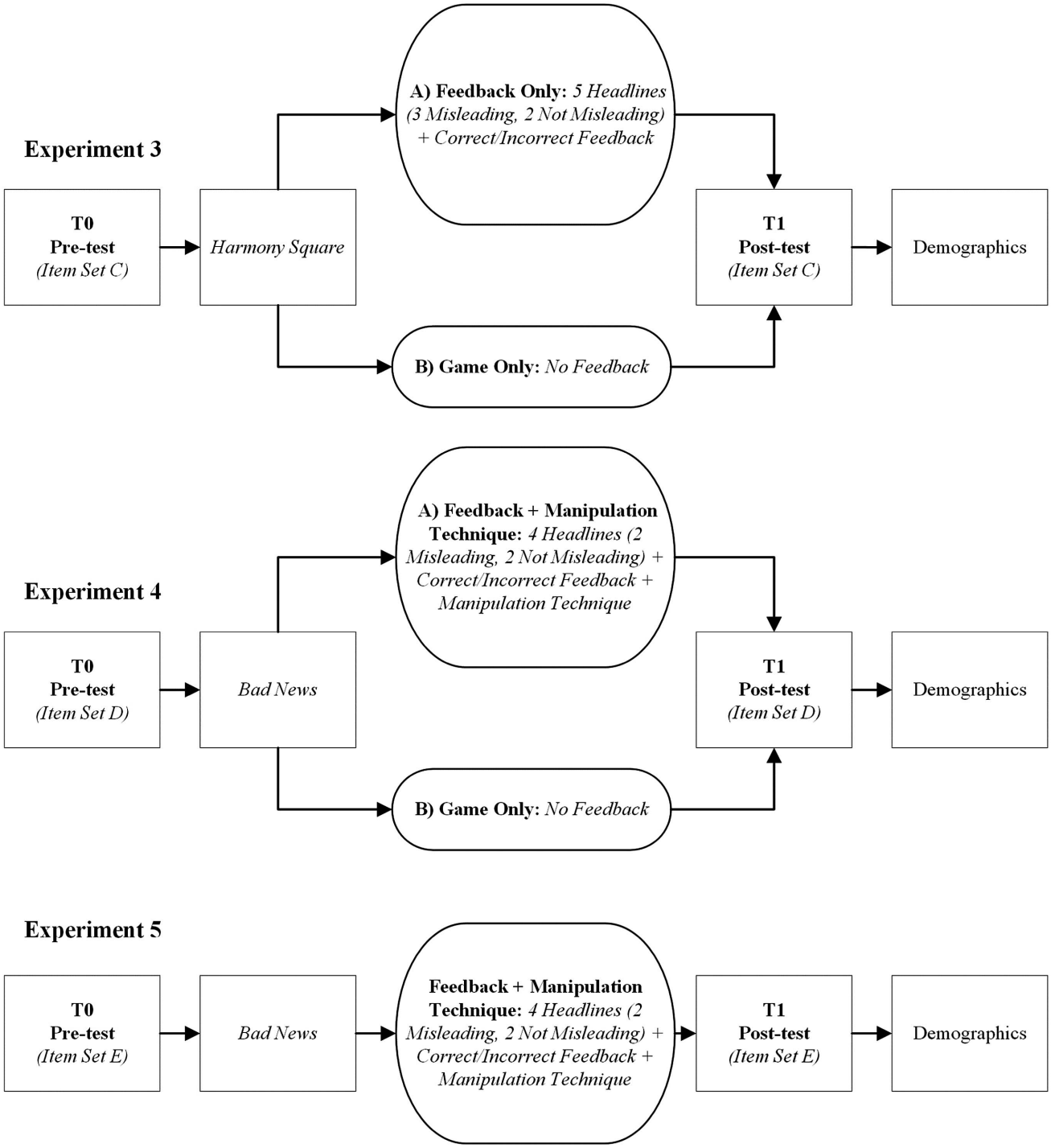
task (consisting of rating the reliability of a series of tweets or headlines on a 1–7-point Likert scale, similar to Experiments 1 and 2) both at the start and the end of each game, as well as a series of demographic questions. Responses from participants who gave informed consent and who completed both the pre- and postgame surveys were recorded. As both games have an organic player base, we were able to collect a substantial amount of data over time. We then let the games collect data both with and without a feedback exercise activated at the end of the game (for Experiments 5 and 6, we only collected data with feedback). Doing so allowed us to compare players' performance on the item rating tasks with and without feedback. These studies were approved by the Cambridge Psychology Research Ethics Committee (PRE.2022.121). We discuss the specifics for each experiment below. See Figure 5 for a flowchart of the study designs. See Table 1 for study details.

Experiment 3

The item sets used in Experiments 1, 2, and 4 have unbalanced misinformation–real news ratios. We therefore administered a balanced eight-item rating task consisting of four tweets containing misinformation and four real news tweets, both at the start and at the end of the *Harmony Square* game (item set C); see Supplemental Table S9 for the scale properties. The participants were asked to

Figure 5

Flowchart Depicting the Procedure of Experiments 3, 4, and 5



Note. Design for Experiment 6, not shown here, is identical to Experiment 5 but with item set C (used in Experiment 3).

indicate their agreement with the statement "this post is reliable" for each tweet, with 1 being *not at all* and 7 being *very*, identical to Study 1 from Basol et al. (2021). The tweets were stripped of all source and social information to avoid these potential confounds. See Figure 6 for examples of a real news and misinformation tweet as shown in the game; see Supplemental Table S24 for the item wordings. In addition, the participants were asked (at the end of the game) to indicate their age group, gender, education level, country/

Table 1
Experiments 3–6: Overview of Data Collection

Experiment	Game	Item set	Item set detail	Purpose	Data collection date	Sample size
3	*Harmony Square*	C	Four real and four misinformation items, no source information	Test feedback versus no feedback	1/10/23–14/2/2023	559
4	*Bad News*	D	Two real and six misinformation items, with source information	Test feedback versus no feedback	8/4/2023–27/5/2023	2,558
5	*Bad News*	E	Four real and four misinformation headlines (MIST-8), no source information	Test feedback tool on validated and balanced item set	25/7/2023–18/8/2023	419
6	*Bad News*	C	Four real and four misinformation items, no source information	Preliminary comparison of *Bad News* and *Harmony Square*	1/1/2024–24/1/2024	882

Note. MIST-8 = 8-item Misinformation Susceptibility Test.

region of origin, and political ideology. The participants who indicated being under 18 years old were excluded as per our ethics approval. In addition, we excluded participants not from the United States.

The feedback tool was administered at the end of the *Harmony Square* game, just before the posttest. Here, the participants were shown five news headlines (three being misleading, two not; the headlines were not the same as those from the item rating task) and were asked to indicate whether they believed each of them was misleading or not misleading. Based on whether they got the answer right, they were then given positive or negative feedback for each headline. For example, one (misleading) headline read: "Chipping imminent? Microchip company sets up office in 'Big Pharma's hometown.'" If players indicated that they thought this was misleading, the game told them "Yep that's right! This is misleading. The headline implies a connection that isn't there by alleging a hidden conspiracy." If they indicated that they found this headline not misleading, the game told them "Nope. That one's clearly misleading. The headline implies a connection that isn't there by alleging a hidden conspiracy." This was done for each of the five headlines. If players indicated that a real news headline was misleading, the game warned them not to be overly skeptical and that "sometimes a headline is just a headline." See Supplemental Tables S25 for the wordings for each feedback headline.

The *Harmony Square* game is a live intervention and constantly collecting data (in the current version of the game, the feedback tool is activated), so our data collection periods with and without the feedback tool spanned a roughly similar amount of time (in an attempt to avoid substantial differences in sample size, we followed the same procedure in Experiment 4). Specifically, we collected data without feedback between October 1 and November 24, 2022, and with feedback between November 25, 2022, and February 14, 2023. In total, we collected 257 valid responses without the feedback tool and 302 valid responses with the feedback tool, for a total of $N = 559$ responses. See Supplemental Table S4 for the sample descriptives.

Experiment 4

We administered an eight-item rating task consisting rating the reliability (1 [*not at all*] and 7 [*very reliable*]) of two real news tweets (the first being a headline by *The Guardian* about Amazon, Apple, and Google being the biggest brands in the world and the second being a tweet by *Psychology Today* saying that physical exercise keeps your brain in good shape) and six misinformation tweets, one for each manipulation technique learned about in the *Bad News* game (item set D); see Supplemental Table S9 for the scale properties. This procedure is identical to the one used in Study 1 from Roozenbeek, Traberg, and van der Linden (2022). See Supplemental Table S24 for the item wordings. In addition, the participants were asked for their age group, gender, political ideology, and trust in the mainstream media (from 1 [*not at all*] to 5 [*a lot*]). Finally, we implemented an attention check in the game, which we used as an exclusion criterion (participants were asked what game they were currently playing).

The feedback tool differed from the one from Experiment 3 in several ways. First, although the headlines themselves were identical,

Figure 6
Experiment 3: Examples of a Real News (Left) and Misinformation (Right) Survey Item Implemented in the Harmony Square Game

Note. See the online article for the color version of this figure.

the exercise consisted of four instead of five headlines (meaning that one headline that was used in Experiment 3 was not used here; see Supplemental Table S25). Second, after indicating whether they found a headline misleading or not if they selected misleading, participants were asked "What makes you think this headline is misleading?" and could indicate which manipulation technique they believed was being used in it (with response options including "it evokes strong negative emotions" or "this is just a headline; I see nothing wrong with it"). Based on their answers, the participants (like in Experiment 3) were then given positive or negative feedback. For instance, if they correctly identified the "Big Pharma" headline from above as misleading but indicated the wrong manipulation technique (polarization instead of conspiratorial thinking), the game may say something like "That's true but not quite what I was looking for. To be more precise: it's conspiratorial. The headline implies that the microchip company and 'Big Pharma' are concocting an evil plot together." See Figure 7 for an example of how the feedback tool looked in the *Bad News* game. See Supplemental Table S25 for the wordings for each feedback headline.

Data were collected between April 8 and 29 (with feedback) and between April 30 and May 27, 2023 (without feedback). In total, we collected 1,342 participants without feedback (960 valid paired pre- and postresponses, as some responses were incomplete) and 1,216 participants with feedback (930 valid paired pre- and postresponses), for a total sample size of 2,558 participants (1,890 valid pre- and postresponses). See Supplemental Table S5 for the sample descriptives.

Experiment 5

Experiment 5 was identical to Experiment 4 except for the item set used in the pre- and posttests. Here, we used the eight-item Misinformation Susceptibility Test, a psychometrically validated test of news veracity discernment (item set E; Maertens et al., 2023). The Misinformation Susceptibility Test–8 (MIST-8) consists of four true and four false headlines (without source information), with participants indicating how reliable they believed each headline to be on a 7-point Likert scale.[7] See Supplemental Table S24 for the item wordings.

A previous study (Maertens et al., 2023) implemented the MIST in the *Bad News* game (without feedback) and found no significant pre- and postdifferences for veracity discernment. We therefore only collected data with the feedback tool active for this experiment to see if the feedback tool can contribute to boosting game players' accuracy in identifying true and false news headlines. Data were collected between July 25 and August 18, 2023. In total, we collected 419 valid pre- and postresponses. See Supplemental Table S6 for the sample descriptives.

Experiment 6

The goal of Experiment 6 ($N = 882$) was to be able to offer a preliminary comparison between the *Harmony Square* and *Bad News* games in terms of participants' performance on the same item rating task. This study was identical to Experiment 5 in every way, except we used item set C (from Experiment 3) instead of the MIST as the item set for the pre- and postsurvey in *Bad News*. For the sake of brevity, we report only the main results in Table 1. For a detailed description of the study design, sample, and results and a discussion, see Supplemental Material K.

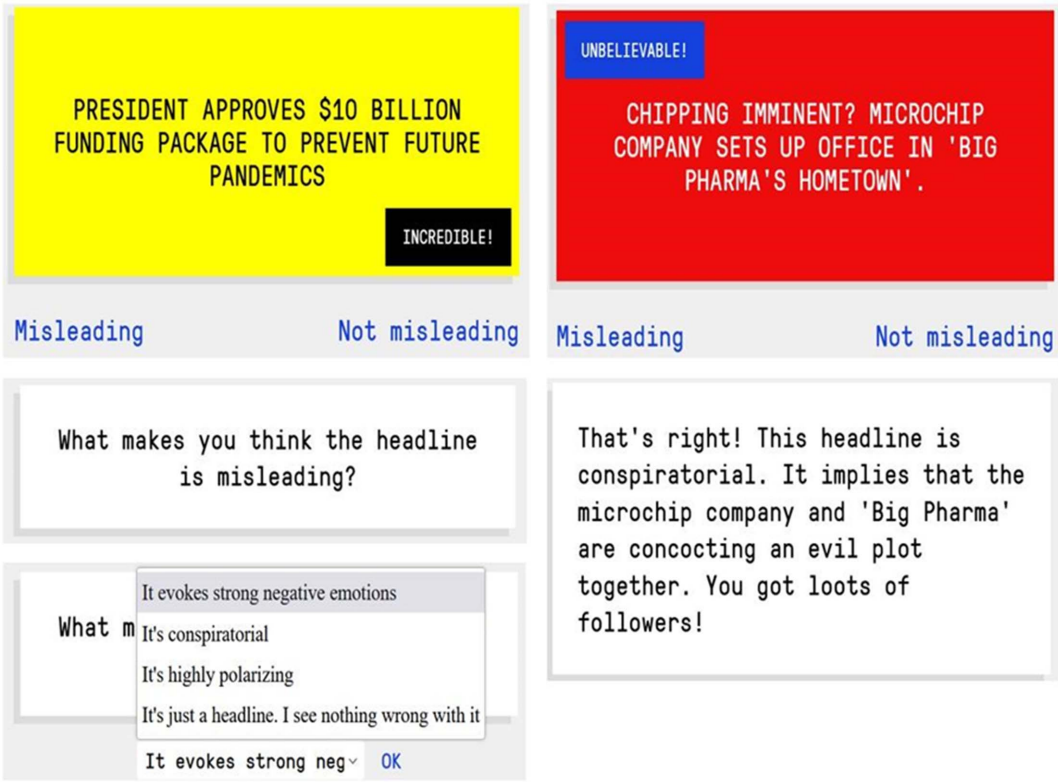
Results

We present the results of our hypothesis tests for Experiments 3–6 in the same table, following the analysis plan used in previous studies with similar or identical designs (Basol et al., 2021; Roozenbeek & van der Linden, 2019; see Table 2). Supplementary Bayesian analyses show that the below results are robust, with all Bayes factors for veracity discernment (Hypothesis 2c) being larger than 813 (i.e., strong evidence against the null hypothesis of no pre- and postdifference in discernment) when the feedback tool is active; see Supplemental Table S22 for a summary. See Supplemental Table S16–S18 for item-level descriptive statistics and Supplemental Tables S19–S21 for item-level t tests (frequentist and Bayesian). See Supplemental Material K for a detailed discussion of the results for Experiment 6.

Table 1 shows that, without feedback, both games reduce the perceived reliability of misinformation as well as real news, indicating support for Hypothesis 1a but not Hypothesis 1b. Hypothesis 1c (discernment) is (weakly) supported for the *Bad News* game (Experiment 4; we note that this hypothesis is also not supported under a Bayesian framework, $BF_{10} = 0.274$; see Supplemental Table S22), but not for *Harmony Square* (Experiment 3). Item-level analyses (see Supplemental Table S20) show that for *Bad News*, the negative effect of the intervention on real news is driven by one of the two real headlines (the one by *Psychology Today* about physical exercise keeping your brain in good shape, $p < .001$), whereas the other real headline (from the *Guardian*, about Apple, Google, and Amazon) showed no significant reduction in reliability postgameplay ($p = .068$). It is possible that a source familiarity effect is at play here: As noted by Modirrousta-Galian and Higham (2023), people do not evaluate uncontroversial true information any differently after playing an inoculation game but rather become more skeptical of *ambiguous* (true) information. It is possible that game players were more familiar with *The Guardian* than *Psychology Today* as a source of reliable information and therefore did not become as uncertain about their assessment of the headline's reliability. Experiments 3, 5, and 6 did not include source information in the headline sets. For Experiment 3, this may have meant that the participants were less certain about the real items' veracity in the pretest (as they were more ambiguous), which then led to reduced reliability postgameplay. Overall, the findings from Experiments 3 and 4 contrast with those from Experiments 1 and 2, in that we did not find a significant reduction in the perceived reliability of real news in the latter two studies; this is in line with other studies that report inconsistent findings in this regard. For example, Roozenbeek and van der Linden (2019) find no change in real news item ratings after playing *Bad News* without feedback, whereas some other studies do (e.g., Graham et al., 2023; Roozenbeek et al., 2021). It is possible that variations in study design or item sets underlie this inconsistency.

[7] This procedure deviates from the original MIST, which has a binary response mode (real or fake). Roozenbeek, Maertens, et al. (2022) tested whether varying the response mode or the question framing (e.g., reliability, trustworthiness, real vs. fake) yields substantially different response patterns for the MIST. They found that even though there was some variation, the MIST retains decent psychometric properties if a different response mode or question framing is used.

Figure 7
Experiment 4: Screenshots From the Feedback Tool in the Bad News Game



Note. See the online article for the color version of this figure.

However, when the feedback tool is active, all four studies show that participants become significantly better at correctly identifying real news postgameplay (in support of Hypothesis 2b), leading to substantially better discernment (supporting Hypothesis 2c). That said, although the game with feedback significantly reduced average reliability ratings of misinformation, we find no support that feedback improves accuracy in detecting misinformation compared with no feedback (meaning mixed results for Hypothesis 2a); rather, participants in the feedback conditions still rated misinformation as significantly less reliable postgameplay but at a descriptively smaller effect size than if no feedback is provided. This is in line with the findings from Experiment 2, where the feedback + manipulation technique information feedback exercise did not yield a significant improvement in the detection of misinformation compared with the control group. Interestingly, we also find differences between the *Harmony Square* and *Bad News* games when testing the feedback tool on the same item set; while discernment is highly significant in both Experiments 3 and 6, the effect size for Experiment 3 is descriptively higher than for Experiment 6 ($d = 0.567$ vs. $d = 0.334$). This may be due to differences in either the games themselves or the implementation of the feedback task; see Supplemental Material K for a discussion.

We also conducted a supplementary analysis using signal detection theory, following the method used by Modirrousta-Galian and Higham (2023); see our OSF page (https://osf.io/cnr6t/) for the analysis scripts. We find that in all four experiments, the feedback condition improved the AUC or area under the curve (a measure of discrimination/veracity discernment) compared with when no feedback tool was active; this effect was significant in Experiments 3 ($p < .001$, $d = .388$), 4 ($p = .037$, $d = .063$), and 6 ($p < .001$, $d = .185$), but not in Experiment 5 ($p = .134$, $d = .060$); the latter effect trends in the right direction and may thus be significant with a higher sample size. Moreover, in Experiments 3 and 4, response bias is reduced (and not present in Experiment 5 and negative in Experiment 6) with the feedback tool active. See Supplemental Table S23 and Figure S12 for details.

Overall, Experiments 3–6 yield similar results to Experiments 1 and 2: in both the *Bad News* and *Harmony Square* games, implementing a feedback tool at the end of the game substantially and significantly boosts both accuracy in identifying real news and discernment between real news and misinformation, while mitigating response bias. However, unlike in Experiments 1 and 2, we note that we were unable to control for order effects (in the sense that headlines/items were all displayed in the same order for each participant; this is a feature of the data collection tool in the Bad News and Harmony Square games, which does not allow for randomized item presentation order).

General Discussion

Across six separate experiments (two preregistered lab studies and four in-game field studies), we find that inoculation games such as *Bad News* and *Harmony Square* benefit from implementing feedback exercises at the end of the game. We show that administering a task where people evaluate a series of headlines or social media posts and

Table 2

Experiments 3–6: Paired-Sample Students for Reliability Ratings of Misinformation, Real News, and Discernment, Pre- and Postgameplay, Without and With Feedback

Experiment	H	Condition		*t*	*df*	*p*	M_{diff}	Cohen's *d*	95% CI
Harmony Square (Exp. 3)		Without feedback							
	H1a	Misinformation (post)	Misinformation (pre)	−7.337	256	<.001*	−.378	−0.458	[−.586, −.329]
	H1b	Real news (post)	Real news (pre)	−5.126	256	<.001*	−.361	−0.320	[−.445, −.194]
	H1c	Discernment (post)	Discernment (pre)	.796	256	.796	.018	0.016	[−.106, .138]
		With feedback							
	H2a	Misinformation (post)	Misinformation (pre)	−2.72	301	.007*	−.192	−0.157	[−.270, −.043]
	H2b	Real news (post)	Real news (pre)	9.896	301	<.001*	.769	0.569	[.447, .691]
	H2c	Discernment (post)	Discernment (pre)	9.847	301	<.001*	.961	0.567	[.445, .688]
Bad News (Exp. 4)		Without feedback							
	H1a	Misinformation (post)	Misinformation (pre)	−9.75	959	<.001*	−.370	−0.315	[.379, −.250]
	H1b	Real news (post)	Real news (pre)	−5.88	959	<.001*	−.272	−0.190	[−.253, −.126]
	H1c	Discernment (post)	Discernment (pre)	2.015	959	.049*	.098	0.065	[.002, .128]
		With feedback							
	H2a	Misinformation (post)	Misinformation (pre)	−7.165	929	<.001*	−.259	−0.235	[−.300, −.170]
	H2b	Real news (post)	Real news (pre)	4.500	929	<.001*	.196	0.148	[.083, .212]
	H2c	Discernment (post)	Discernment (pre)	9.007	929	<.001*	.455	0.295	[.230, .361]
Bad News (Exp. 5)		With feedback							
	H2a	Misinformation (post)	Misinformation (pre)	−2.976	418	.003*	−.180	−0.145	[−.242, −.049]
	H2b	Real news (post)	Real news (pre)	3.49	418	<.001*	.190	0.171	[.074, .267]
	H2c	Discernment (post)	Discernment (pre)	5.421	418	<.001*	.370	0.265	[.167, .362]
Bad News (Exp. 6)		With feedback							
	H2a	Misinformation (post)	Misinformation (pre)	−3.600	881	<.001*	−.155	0.121	[−.187, −.055]
	H2b	Real news (post)	Real news (pre)	7.331	881	<.001*	.358	0.247	[.180, .314]
	H2c	Discernment (post)	Discernment (pre)	9.913	881	<.001*	.512	0.334	[.266, .402]

Note. Column H lists the hypotheses; green color indicates that the hypothesis was supported, red color indicates that it was not supported; black indicates mixed evidence. See also Supplemental Table S22. Exp. = Experiment. See the online article for the color version of this table.
* $p < .05$.

receive feedback based on how they did significantly increases their ability to discern misinformation from real news, compared with a version of the game where no feedback is provided. In addition, we show that this improved performance persists over time, with people who received feedback evaluating both misinformation and real news more accurately 1 week postgameplay than people who did not. Using an ordinal probit regression in Experiments 1 and 2, we were able to show not only that *Bad News* (and other inoculation games) shift responses on item rating tasks but also how homogenous this shift is. Importantly, we find that introducing a feedback mechanism boosts overall item rating task performance, meaning that not only the average but also the *distribution* of the responses is shifted. Moreover, these results were reinforced and replicated in Experiments 3–5 using signal detection theory, boosting the area under the curve (discrimination/veracity discernment) and significantly reducing or mitigating response bias.

We also report preliminary findings that feedback exercises increase the longevity of the intervention, as improved performance was detected 1 week postgameplay in Experiments 1 and 2. It is possible that this type of feedback exercise helps strengthen the memory of the lessons from the intervention, boosting both performance and longevity (Bird et al., 2015; Maertens et al., 2024). Providing participants with an additional learning opportunity immediately after the intervention may be effective at increasing the long-term effectiveness of inoculation through memory strengthening and the integration of what is learned in associative memory networks (Pfau et al., 2005). In line with other recent findings (Capewell et al., 2024; Maertens et al., 2024), we therefore suggest that feedback and practice are likely critical components of successful learning-based misinformation interventions (not just inoculation games) and "boosting" interventions more generally (Hertwig & Grüne-Yanoff, 2017).

Interestingly, as shown in Experiments 3–6, the increase in discernment is driven primarily by an increase in people's accuracy in detecting real news: With feedback, players become *less* skeptical of real news after playing an inoculation game. This is good news, as this reduces concerns over misinformation interventions potentially causing undue skepticism of true information (Hameleers, 2023; Modirrousta-Galian & Higham, 2023). That said, it is important to contextualize the nuances of any real news effect. It has been argued that an intervention yielding a more conservative response bias (i.e., people becoming more skeptical of real as well as false information) could have "devastating consequences" such as people losing trust in vaccines (Modirrousta-Galian & Higham, 2023, p. 24). We believe this assertion to be hyperbolic for several reasons. First, there is no evidence for this claim, and it seems unlikely that merely playing a humorous game or reading a set of benign media literacy tips (Hameleers, 2023) could lead to people losing their belief in verifiable facts. On the contrary, research finds that although people might question specific (true) headlines after some interventions, there was no reduction in overall trust in media or institutions (Hoes et al., 2023). Effectively, the argument boils down to the idea that the mere act of providing people with information about how they might be manipulated online can cause harm. This is reminiscent of the "backfire effect" of debunking misinformation, which despite initial concerns (Nyhan & Reifler, 2010) was later shown to not be reliably observed and mostly due to a design artifact (Ecker et al., 2020; Swire-Thompson et al., 2020, 2022; Wood & Porter, 2019).

Second, as shown in the present study as well as other studies (Hameleers, 2023; Modirrousta-Galian & Higham, 2023), there is no evidence that misinformation interventions (including inoculation games) lead to *distrust* of real news or true information writ large, only *reduced* trust in some (but not all) true headlines. As is the case in all these studies (which universally show that people continue to rate real news as more reliable/true than not), if people evaluate a real news headline as highly reliable (e.g., 6.5 out of 7) before an intervention, and as slightly less reliable (e.g., 6 out of 7) afterward, does this demonstrate that people have come to distrust this headline? Perhaps not: 6 out of 7 is still well above the midpoint of the scale (and in the *very reliable* range), and so while one could argue that this effect is unintended and a suboptimal outcome of the intervention in terms of item rating task performance (more on this below), the claim that an intervention can lead to a generalized disbelief in such a headline (let alone true *information* in general) is contentious.

Another potential issue is the relevance of people's belief in real news relative to belief in fake news (or misinformation) in terms of its consequences for society. For instance, Loomba et al. (2021, 2023) found that, in the United Kingdom at least, susceptibility to fake news was independently (and negatively) associated with regional COVID-19 vaccine uptake, whereas belief in real news was not (and neither was veracity discernment, although this measure did trend in the right direction). In other words, while people's belief in fake news can be translated to real-world behavior (in this case vaccine uptake), this is less clear for real news; to what extent real news belief thus relates to belief in true information in a general sense, and what misinformation interventions achieve when they decrease or increase belief in real news on item rating tasks, therefore requires further attention.

Finally, whether the real news effect is observed at all is strongly related to intervention design considerations and not a necessary or even common result of inoculation games (e.g., see Barzilai et al., 2023; Lees et al., 2023; Ma et al., 2023, whose games showed good discernment effects) and, in turn, media literacy tips that can also cause real news "skepticism" (Hameleers, 2023; Hoes et al., 2023). As we have shown here, the real news effect is easily mitigated and even reversed entirely using a simple tweak to the intervention in the form of a short feedback exercise. If this effect is so easy to reverse, it is highly unlikely that there is something about misinformation interventions (or inoculation games specifically) that could lead to generalized distrust or skepticism of true information. Instead, the real news effect discussion pertains to suboptimal performance on an outcome measure (in this case an item rating task) and how to optimize learning environments. While an important component of measuring an intervention's efficacy, this performance is relatively easy to optimize because the games are "living" interventions that can be updated. We therefore urge the field to look beyond item rating task performance as a narrow measure of intervention efficacy and also take into consideration other factors such as an intervention's entertainment value and potential scalability and adaptability as equally important measures of effectiveness.

Constraints on Generality

In Experiments 1, 4, 5, and 6, our sample consisted of participants from all over the world (albeit with a heavy focus on Western countries, particularly the United States), whereas Experiment 2 only included participants from the United Kingdom, and Experiment 3 was limited to participants from the United States. Considering the mixed results of studies that tested the efficacy of inoculation games in non-Western contexts (Harjani et al., 2023; Iyengar et al., 2022; Saleh et al., 2023), we cannot assume that our results generalize globally, and we therefore urge caution with the interpretation of our findings. As we were able to conceptually replicate the effect of feedback in all six experiments, we assume that there is at least some level of cross-cultural validity. However, in Experiment 1, the effect of *Bad News* (control group) only resulted in a small (but nonetheless significant) difference in reliability ratings when comparing the baseline measure with T2. Furthermore, the feedback in Experiment 1 had a stronger effect than any feedback condition in Experiment 2. This suggests that in a diverse international sample with heterogeneous language skills/familiarity with Western news content, feedback might be particularly beneficial.

Moreover, in Experiment 2, we observed that the feedback + manipulation technique condition performed worse than the two other feedback conditions, particularly as there was no significant effect (compared with the control group) of this type of feedback exercise on misinformation ratings. This is surprising, as the group only received additional information on top of the feedback, and this type of feedback worked well in Experiments 4 and 5 (albeit descriptively less well than the "feedback only" type, cf. Experiments 3 and 6). It is possible that this information distracted participants or resulted in participants trying to guess the strategies used and not focus on rating the reliability of the items, which was our main performance measure.

Another constraint is that we could not test the effect of the feedback exercises in isolation, without playing an inoculation game. When looking only at participants' performance on item rating tasks, it is possible that some kind of feedback exercise alone might yield a boost in the ability to discern misinformation from real news, although previous research has shown mixed results (Epstein et al., 2021; Modirrousta-Galian et al., 2022). However, as discussed above, we stress the importance of looking beyond item rating task performance as a measure for whether misinformation interventions "work" (Guay et al., 2023): A hypothetical intervention that does nothing but ask participants to evaluate a few headlines and receive feedback is likely not very scalable, as there is little motivation for people to participate nor would it be useful in, for example, educational settings. Inoculation games contain a storyline, may (or may not) make use of humor, include reward and punishment mechanics, use perspective taking, and feature additional information about how misinformation works. While certainly not everyone enjoys *Bad News* or *Harmony Square*, the games' organic player base and widespread use in education indicates that there is a considerable degree of voluntary participation, which we argue is a measure of the intervention "working" alongside optimizing item rating task performance.

Conclusion

Across six separate experiments, we show that inoculation games benefit from implementing feedback exercises where players are briefed about their performance on item rating tasks (in this case tasks that seek to assess people's ability to recognize misinformation). This simple tweak to the interventions results in participants rating misinformation as substantially less reliable after gameplay compared with before while eliminating and even reversing potentially

unwanted skepticism of real news (or non-misinformation), resulting in increased veracity discernment. In addition, we find preliminary evidence that feedback exercises may boost the longevity of such interventions, as we find stronger inoculation effects 1 week postgameplay if feedback is provided compared with playing an inoculation game without feedback. These effects are robust using numerous methods of analyses (including signal detection theory), across different item sets, and for different types of feedback exercise. Creators of gamified interventions (not only about misinformation but about a variety of topics) may consider these findings during the design process. For instance, developers may include a short feedback exercise at the end of their game to test players' skills in a particular outcome measure (such as identifying manipulation techniques in news headlines or social media content). We have now done so in the *Bad News, Harmony Square*, and *Go Viral!* (https://www.goviralgame.com) games.

References

Bandura, A. (1997). *Self-efficacy: The exercise of control*. WH Freeman/TimesBooks/Henry Holt.

Barzilai, S., Mor-Hagani, S., Abed, F., Tal-Savir, D., Goldik, N., Talmon, I., & Davidow, O. (2023). Misinformation is contagious: Middle school students learn how to evaluate and share information responsibly through a digital game. *Computers & Education*, *202*, Article 104832. https://doi.org/10.1016/j.compedu.2023.104832

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 & Society*, *8*(1). https://doi.org/10.1177/20539517211013868

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

Beierlein, C., Kemper, C. J., Kovaleva, A., & Rammstedt, B. (2012). *Kurzskala zur Messung des zwischenmenschlichen Vertrauens: Die Kurzskala Interpersonales Vertrauen (KUSIV3)* (GESIS working papers 2012|22). GESIS. https://pub.uni-bielefeld.de/record/2575625#apa

Bertolotti, M., & Catellani, P. (2023). Counterfactual thinking as a prebunking strategy to contrast misinformation on COVID-19. *Journal of Experimental Social Psychology*, *104*, Article 104404. https://doi.org/10.1016/j.jesp.2022.104404

Betsch, C. (2004). Präferenz für Intuition und Deliberation (PID) [Preference for intuition and deliberation (PID)]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *25*(4), 179–197. https://doi.org/10.1024/0170-1789.25.4.179

Bird, C. M., Keidel, J. L., Ing, L. P., Horner, A. J., & Burgess, N. (2015). Consolidation of complex events via reinstatement in posterior cingulate cortex. *The Journal of Neuroscience*, *35*(43), 14426–14434. https://doi.org/10.1523/JNEUROSCI.1774-15.2015

Bong, M. (2002). Predictive utility of subject-, task-, and problem-specific self-efficacy judgments for immediate and delayed academic performances. *Journal of Experimental Education*, *70*(2), 133–162. https://doi.org/10.1080/00220970209599503

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), Article 395. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

Capewell, G., Maertens, R., Remshard, M., Compton, J., Van der Linden, S., Lewandowsky, S., & Roozenbeek, J. (2024). *Misinformation interventions decay rapidly without an immediate post-test*. PsyArXiv. https://doi.org/10.31234/osf.io/93ujx

Chan, J. C. Y., & Lam, S. (2010). Effects of different evaluative feedback on students' self-efficacy in learning. *Instructional Science*, *38*(1), 37–58. https://doi.org/10.1007/s11251-008-9077-2

Christensen, R. H. B. (2019). *ordinal: Regression models for ordinal data* [Computer software]. R package published through CRAN (R package Version 2019.12-10). https://cran.r-project.org/package=ordinal

Compton, J. (2013). Inoculation theory. In J. P. Dillard & L. Shen (Eds.), *The SAGE handbook of persuasion: Developments in theory and practice* (pp. 220–236). Sage Publications. https://doi.org/10.4135/9781452218410

Compton, J., Van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), Article e12602. https://doi.org/10.1111/spc3.12602

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. (2023). The Cranky Uncle game—Combining humor and gamification to build student resilience against climate misinformation. *Environmental Education Research*, *29*(4), 607–623. https://doi.org/10.1080/13504622.2022.2085671

DROG. (2018). Bad news. [*Information Sheet*]. https://www.getbadnews.com/#intro

Ecker, U. K. H., Lewandowsky, S., & Chadwick, M. (2020). Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, *5*(1), Article 41. https://doi.org/10.1186/s41235-020-00241-6

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School: Misinformation Review*. https://doi.org/10.37016/mr-2020-71

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Graham, M. E., Skov, B., Gilson, Z., Heise, C., Fallow, K. M., Mah, E. Y., & Lindsay, D. S. (2023). Mixed news about the bad news game. *Journal of Cognition*, *6*(1), Article 58. https://doi.org/10.5334/joc.324

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, *94*, 149–165. https://doi.org/10.1016/j.jml.2016.11.001

Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour*, *7*(8), 1231–1233. https://doi.org/10.1038/s41562-023-01667-w

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Gunther, R., Beck, P. A., & Nisbet, E. C. (2019). "Fake news" and the defection of 2012 Obama voters in the 2016 presidential election. *Electoral Studies*, *61*, Article 102030. https://doi.org/10.1016/j.electstud.2019.03.006

Hameleers, M. (2023). The (un)intended consequences of emphasizing the threats of mis- and disinformation. *Media and Communication*, *11*(2). https://doi.org/10.17645/mac.v11i2.6301

Harjani, T., Basol, M., Roozenbeek, J., & van der Linden, S. (2023). Gamified inoculation against misinformation in India: A randomised control trial. *Journal of Trial and Error*, *3*(1), 14–56. https://doi.org/10.36850/e12

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, *12*(6), 973–986. https://doi.org/10.1177/1745691617702496

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2023). *Prominent Misinformation Interventions Reduce Misperceptions but Increase Skepticism*. PsyArxiv. https://doi.org/10.31234/osf.io/zmpdu

Hopp, T. (2022). Fake news self-efficacy, fake news identification, and content sharing on Facebook. *Journal of Information Technology & Politics*, *19*(2), 229–252. https://doi.org/10.1080/19331681.2021.1962778

Iyengar, A., Gupta, P., & Priya, N. (2022). Inoculation against conspiracy theories: A consumer side approach to India's fake news problem. *Applied Cognitive Psychology*, *37*(2), 290–303. https://doi.org/10.1002/acp.3995

Jang, A. I., Nassar, M. R., Dillon, D. G., & Frank, M. J. (2019). Positive reward prediction errors during decision-making strengthen memory encoding. *Nature Human Behaviour*, *3*(7), 719–732. https://doi.org/10.1038/s41562-019-0597-3

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *The Journal of Politics*, *62*(3), 790–816. https://doi.org/10.1111/0022-3816.00033

Lees, J., Banas, J. A., Linvill, D., Meirick, P. C., & Warren, P. (2023). The spot the troll quiz game increases accuracy in discerning between real and inauthentic social media accounts. *PNAS Nexus*, *2*(4), Article pgad094. https://doi.org/10.1093/pnasnexus/pgad094

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

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

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1

Loomba, S., Götz, F. M., Maertens, R., Roozenbeek, J., de Figueiredo, A., & van der Linden, S. (2023). *Ability to detect fake news predicts geographical variation in COVID-19 vaccine uptake*. MedArXiv. https://doi.org/10.1101/2023.05.10.23289764

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*, Article e49255. https://doi.org/10.2196/49255

Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change*, *58*, Article 101964. https://doi.org/10.1016/j.gloenvcha.2019.101964

Ma, J., Chen, Y., Zhu, H., & Gan, Y. (2023). Fighting COVID-19 misinformation through an online game based on the inoculation theory: Analyzing the mediating effects of perceived threat and persuasion knowledge. *International Journal of Environmental Research and Public Health*, *20*(2), Article 980. https://doi.org/10.3390/ijerph20020980

Maertens, R., Götz, F. M., Golino, H., Schneider, C., Roozenbeek, J., Kerr, J., Stieger, S., McClanahan, W. P., Drabot, K., & van der Linden, S. (2023). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods*, *56*, 1863–1899. https://doi.org/10.31234/osf.io/gk68h

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

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*. PsyArxiv. https://doi.org/10.31234/osf.io/6r9as

McGrew, S. (2020). Learning to evaluate: An intervention in civic online reasoning. *Computers & Education*, *145*, Article 103711. https://doi.org/10.1016/j.compedu.2019.103711

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

McGuire, W. J. (1961b). The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion. *Sociometry*, *24*(2), Article 184. https://doi.org/10.2307/2786067

McGuire, W. J. (1964). Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 1, pp. 191–229). Academic Press. https://doi.org/10.1016/S0065-2601(08)60052-0

McGuire, W. J., & Papageorgis, D. (1962). Effectiveness of forewarning in developing resistance to persuasion. *Public Opinion Quarterly*, *26*(1), 24–34. https://doi.org/10.1086/267068

Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General*, *152*(9), 2411–2437. https://doi.org/10.1037/xge0001395

Modirrousta-Galian, A., Higham, P. A., & Seabrooke, T. (2022). *Effects of inductive learning and gamification on news veracity discernment*. PsyArxiv. https://doi.org/10.31234/osf.io/4wfds

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology*, *65*(1), 113–131. https://doi.org/10.1037/0022-3514.65.1.113

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Paciello, M., Corbelli, G., & D'Errico, F. (2023). The role of self-efficacy beliefs in dealing with misinformation among adolescents. *Frontiers in Psychology*, *14*, Article 1155280. https://doi.org/10.3389/fpsyg.2023.1155280

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental

evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

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

Rasmussen, J., Lindekilde, L., & Petersen, M. B. (2022). *Public health communication reduces COVID-19 misinformation sharing and boosts self-efficacy*. PsyArXiv. https://doi.org/10.31234/osf.io/8wdfp

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/10.31234/osf.io/b52um

Roozenbeek, J., Maertens, R., Herzog, S., Geers, M., Kurvers, R., Sultan, M., & van der Linden, S. (2022). Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, *17*(3), 547–573. https://doi.org/10.1017/S1930297500003570

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

Roozenbeek, J., Traberg, C. S., & van der Linden, S. (2022). Technique-based inoculation against real-world misinformation. *Royal Society Open Science*, *9*(5), Article 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: 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. https://doi.org/10.1017/9781009214414

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), Article eabo6254. https://doi.org/10.1126/sciadv.abo6254

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Saleh, N., Makki, F., van der Linden, S., & Roozenbeek, J. (2023). Inoculating against extremist persuasion techniques—Results from a randomised controlled trial in post-conflict areas in Iraq. *Advances in Psychology*, *1*(1), 1–21. https://doi.org/10.56296/aip00005

Scheibenzuber, C., Hofer, S., & Nistor, N. (2021). Designing for fake news literacy training: A problem-based undergraduate online-course. *Computers in Human Behavior*, *121*, Article 106796. https://doi.org/10.1016/j.chb.2021.106796

Segel, L. A., & Bar-Or, R. L. (1999). On the role of feedback in promoting conflicting goals of the adaptive immune system. *Journal of Immunology*, *163*(3), 1342–1349. https://doi.org/10.4049/jimmunol.163.3.1342

Shamon, H., & Berning, C. C. (2020). Attention check items and instructions in online surveys: Boon or bane for data quality? *Survey Research Methods*, *14*(1), 55–77. https://doi.org/10.18148/SRM/2020.V14I1.7374

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 80–95. https://doi.org/10.1037/a0017407

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., & Lazer, D. (2020). Public health and online misinformation: Challenges and recommendations. *Annual Review of Public Health*, *41*(1), 433–451. https://doi.org/10.1146/annurev-publhealth-040119-094127

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

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

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

van der Linden, S. (2024). Countering misinformation through psychological inoculation. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 69, pp. 1–58). Academic Press. https://doi.org/10.1016/bs.aesp.2023.11.001

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

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

Received September 4, 2023
Revision received March 20, 2024
Accepted March 25, 2024 ■