The Harvard Kennedy School Misinformation Review[1]
January 2020, Volume 1, Issue 2

DOI: https://doi.org/10.37016//mr-2020-008
Website: misinforeview.hks.harvard.edu

Research Article

Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures

This study finds that the online "fake news" game, Bad News, can confer psychological resistance against common online misinformation strategies across different cultures. The intervention draws on the theory of psychological inoculation: analogous to the process of medical immunization, we find that "prebunking," or preemptively warning and exposing people to weakened doses of misinformation, can help cultivate "mental antibodies" against fake news. We conclude that social impact games rooted in basic insights from social psychology can boost immunity against misinformation across a variety of cultural, linguistic, and political settings.

Authors: Jon Roozenbeek (1), Sander van der Linden (2), and Thomas Nygren (3)
Affiliations: (1, 2) Department of Psychology, University of Cambridge, UK, (3) Department of Education, Uppsala University
How to cite: Roozenbeek, Jon, van der Linden, Sander, Nygren, Thomas (2020). Prebunking interventions based "inoculation" theory can reduce susceptibility to misinformation across cultures. *The Harvard Kennedy School (HKS) Misinformation Review*, Volume 1, Issue 2
Received: Dec.13, 2019; Accepted: Jan. 24, 2020; Published: Feb. 3rd, 2020

Research questions

- Is it possible to build psychological "immunity" against online misinformation?
- Does *Bad News,* an award-winning fake news game, help people spot misinformation techniques across different cultures?

Essay summary

- We designed an online game in which players enter a fictional social media environment. In the game, the players "walk a mile" in the shoes of a fake news creator. After playing the game, we found that people became less susceptible to future exposure to common misinformation techniques, an approach we call *prebunking*.
- In a cross-cultural comparison conducted in collaboration with the UK Foreign and Commonwealth Office and the Dutch media platform DROG, we tested the effectiveness of this game in 4 languages other than English (German, Greek, Polish, and Swedish).
- We conducted 4 voluntary in-game experiments using a convenience sample for each language version of *Bad News* (*n* = 5,061). We tested people's assessment of the reliability of several fake and "real" (i.e., credible) Twitter posts before and after playing the game.

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University's John F. Kennedy School of Government.*

- We find significant and meaningful reductions in the perceived reliability of manipulative content across all languages, indicating that participants' ability to spot misinformation significantly improved. Relevant demographic variables such as age, gender, education level, and political ideology did not substantially influence the inoculation effect.
- Our real-world intervention shows that social impact games rooted in insights from social psychology can boost psychological immunity against online misinformation across a variety of cultural, linguistic, and political settings.
- Social media companies, governments, and educational institutions could develop similar large-scale "vaccination programs" against misinformation. Such interventions can be directly implemented in educational programs, adapted for use within social media environments, or applied in other issue domains where online misinformation is a threat.
- In contrast to classical "debunking," we recommend that (social media) companies, governmental, and educational institutions also consider *prebunking* (inoculation) as an effective means to combat the spread of online misinformation.

Implications

There is a loud call for educational interventions to help citizens navigate credible, biased, and false information. According to the World Economic Forum (2018), online misinformation is a pervasive global threat. UNESCO underscored how all citizens need better up-to-date knowledge, skills, and attitudes to critically assess online information (Carlsson, 2019).

Common approaches to tackling the problem of online misinformation include developing and improving detection algorithms (Monti et al., 2019), introducing or amending legislation (Human Rights Watch, 2018), developing and improving fact-checking mechanisms (Nyhan & Reifler, 2012), and focusing on media literacy education (Livingstone, 2018). However, such interventions present limitations. In particular, it has been shown that debunking and fact-checking can lack effectiveness because of the continued influence of misinformation: once people are exposed to a falsehood, it is difficult to correct (De keersmaecker & Roets, 2017; Lewandowsky et al., 2012). Overall, there is a lack of evidence-based educational materials to support citizens' attitudes and abilities to resist misinformation (European Union, 2018; Wardle & Derakshan, 2017). Importantly, most research-based educational interventions do not reach beyond the classroom (Lee, 2018).

Inoculation theory is a framework from social psychology that posits that it is possible to *pre-emptively* confer psychological resistance against (malicious) persuasion attempts (Compton, 2013; McGuire & Papageorgis, 1961). This is a fitting analogy, because "fake news" can spread much like a virus (Kucharski, 2016; Vosoughi et al., 2018). In the context of vaccines, the body is exposed to a weakened dose of a pathogen—strong enough to trigger the immune system—but not so strong as to overwhelm the body. The same can be achieved with information by introducing pre-emptive refutations of weakened arguments, which help build cognitive resistance against future persuasion attempts. Meta-analyses have shown that inoculation theory is effective at reducing vulnerability to persuasion (Banas & Rains, 2010).

Traditional inoculation research tends to focus on individual arguments against specific persuasion or disinformation attempts (McGuire, 1961; van der Linden et al., 2017). In a significant departure from this paradigm, we have already shown that, using a game-based intervention, participants can also improve in their latent ability to spot misinformation *techniques* as opposed to just individual instances of misinformation (see Basol et al., 2020; Roozenbeek & van der Linden, 2018, 2019). The main premise of our approach is that fake news stories themselves constantly change and evolve so building immunity against the underlying *tactics* of misinformation is a more durable strategy.

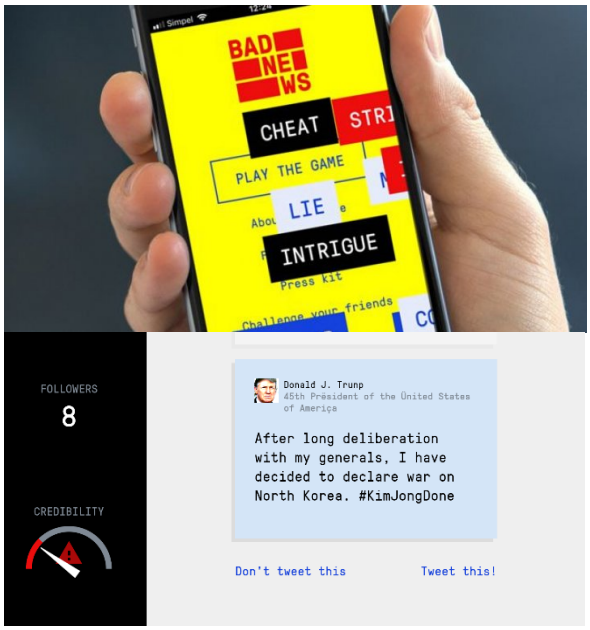

Figure 1. *Bad News landing page and gameplay*

The research we present here focuses on an "active" psychological inoculation intervention called *Bad News*. Bad News is a free choice-based browser game in which players take on the role of fake news creators and learn about 6 common misinformation techniques[2]. In our experiments, we showed that the game is effective at building resistance against misinformation (Basol et al., 2020; Roozenbeek & van der Linden, 2019)[3]. The *Bad News* game was widely covered in the media and has won several awards (BBC News, 2018a; CNN, 2019; Reuters, 2019)[4]. Roozenbeek and van der Linden (2019) co-developed and wrote the game's content and implemented a survey within the game that players can choose to participate in. This initial evaluation, which involved asking participants about the reliability of several "fake" and "credible" Twitter posts pre- and post-gameplay, showed consistent and significant inoculation effects. This evidence was presented as part of the parliamentary Inquiry on Fake News in the UK (van der Linden et al., 2018) and—in light of further randomized trials and empirical testing—subsequently referred to as "one of the most sustainable paths to combating fake news" by the European Commission (Klossa, 2019, p.23).

In a continuation of the *Bad News* project, we collaborated with DROG, a Dutch media literacy platform, as well as the United Kingdom's Foreign and Commonwealth Office (FCO), to translate and adapt *Bad News* in a variety of languages in order to promote media literacy worldwide. The purpose of this collaboration was to scale the Bad News intervention in a variety of languages free for anyone to play and access. Thus far, the intervention has reached about a million people. In the research presented here, we evaluate the cross-cultural effectiveness of the game and find support for its efficacy across a range of different cultural and political contexts.

Practically, our game is used as a digital literacy teaching tool (information for educators is available on the game's website). Importantly, game-based interventions are scalable (i.e. they can reach millions of people worldwide) and offer potential for combating misinformation in other domains as well. For example, we have developed an intervention in collaboration with WhatsApp called "*Join this Group*". The game inoculates players against misinformation frequently encountered within the context of direct messaging apps, which are a growing problem in countries such as India, Mexico, and Brazil (Phartiyal et al., 2018; Roozenbeek et al., 2019). Similarly, people can be inoculated against the techniques used in extremist online recruitment by preemptively warning and exposing them to weakened versions of these techniques. In collaboration with the Lebanese Behavioral Insights Unit (Nudge Lebanon) we have developed a game-based intervention called *Radicalize*, which we recently presented at the United Nations (UNITAR, 2019).

However, it is important to note that digital literacy is complex (Nygren, 2019) and inoculation interventions on their own do not offer a silver bullet to all of the challenges of navigating the post-truth information landscape. Future research may investigate how inoculation interventions can be combined with other interventions to support critical thinking (Lutzke et al., 2019), evaluations of authentic news

[2] The game is online and freely available for anyone to play at www.getbadnews.com

[3] Please see the Method section of this paper to read more about the game.

[4] For example, the *Bad News* game was recently awarded the 'Trust' Prize from the Royal Holland Society of Sciences.

feeds (Nygren et al., 2019), and civic online reasoning (McGrew, 2020). Nonetheless, our results show that this real-world intervention can significantly boost people's ability to recognize online misinformation and improve their resistance against manipulation attempts in a variety of cultural, linguistic, and political settings.

Findings

Finding 1: An online social impact game, Bad News*, is effective at conferring resistance against misinformation techniques in Sweden.*

As a first step, we developed and evaluated the Swedish-language version of the *Bad News* game[5] using an in-game survey as a cross-cultural pilot test and a direct replication of the English version. Accordingly, the measures and methodology were therefore kept similar to the initial evaluation of the game by Roozenbeek and van der Linden (2019) with three fake news Twitter posts (making use of the impersonation, conspiracy, and discrediting techniques) and 3 "credible" news control items that did not contain any deception techniques ("Brexit, the United Kingdom's exit from the European Union, will happen in 2019", "US President Donald Trump wants to build a wall between the US and Mexico" and "Only one in ten students can distinguish between real news and advertisement"[6]). We showed participants (n = 134 to n = 379 for the full pre-post survey) 3 fake and 3 credible (control) Twitter posts before and after playing the game and asked them to rate their reliability on a 1-7 Likert scale (see the "Methods" section for more details). We hypothesized that playing the game would significantly reduce the perceived reliability of fake (but not credible) items.

We find that playing *Bad News* significantly reduces participants' susceptibility to simulated fake Twitter posts for impersonation ($t(173) = -2.37$, $p = 0.018$, $d = 0.18$), conspiracy ($t(378) = -3.22$, $p = 0.001$, $d = 0.17$) and discrediting ($t(375) = -2.74$, $p = 0.001$, $d = 0.15$). Averaged across all fake items, the mean reduction in reliability judgments is also significant ($t(172) = -3.38$, $p = 0.001$, $d = 0.24$). On the other hand, people did not reduce or change their ratings of the credible news control items (for which we hypothesized no significant change), including the survey items about Brexit ($t(133) = -0.88$, $p = 0.38$), Donald Trump ($t(379) = -0.28$, $p = 0.78$), and media literacy ($t(243) = -0.35$, $p = 0.72$). These results are consistent with earlier findings by Roozenbeek & van der Linden (2019) and are visualized in Figure 2 below.

[5] This game can be played for free in any browser at www.badnewsgame.se

[6] The full list of survey items can be found in Table 2 in the appendix.

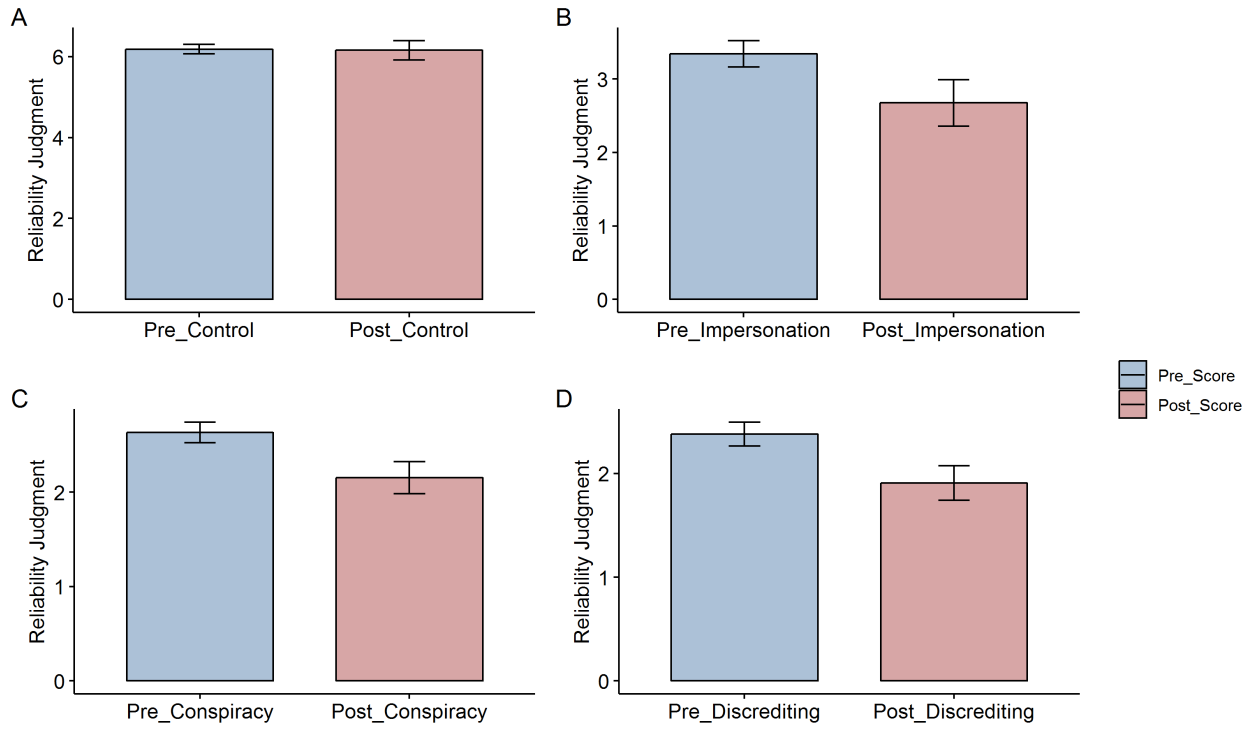

Figure 2. *Perceived reliability judgments for "real" control items (averaged, panel A), impersonation (panel B), conspiracy (panel C), and discrediting (Panel D), for the Swedish version of the game. The error bars represent 95% confidence intervals. The figure shows that the credible news items were perceived as equally reliable before and after playing, whereas the "fake" items are seen as significantly less reliable. All mean pre and post scores can be found in Table 3 in the appendix.*

Finding 2: In a cross-cultural sample, the Bad News game also improves people's ability to recognize misinformation strategies in Germany, Greece, and Poland.

The Swedish pilot study demonstrated the potential cross-cultural effectiveness of the *Bad News* game. The survey items were subsequently extended and standardized across languages to include an item for each of the six manipulation techniques people learn about in the game (impersonation, conspiracy, emotion, polarization, discrediting, and trolling; see Roozenbeek & van der Linden (2019) for a more thorough explanation of these techniques). To corroborate these initial results, we also implemented pre-post surveys in the game's German, Greek and Polish versions[7]. As such, the exact same six items were administered pre- and post-gameplay in Germany (n = 2,038), Greece (n = 1,518), and Poland (n = 1,332) along with standard socio-demographic questions (age, gender, education, and political ideology).

Following Roozenbeek and van der Linden (2019) and Basol et al. (2020), we created a single fake news scale (averaged across all items) to reduce multiple testing (though item-level results are also presented by country in Tables 4, 5, and 6 in the appendix). This allowed us to compare participants' performance across the three countries. Overall, the aggregate inoculation effect across countries is significant (M_{pre} = 2.50, M_{post} = 2.09, M_{diff} = -0.41, [95% CI -0.38, -0.44], (t(4887) = -28.47, p < 0.01, d = 0.37), indicating that the game effectively inoculates players against misinformation techniques across a variety of cultural and linguistic settings[8]. The observed effect-size can be described as close to "medium" or "moderate" and is very similar to the English version of *Bad News* (Roozenbeek & van der Linden, 2019).

[7] These versions can be played at www.getbadnews.de, www.getbadnews.gr, and www.getbadnews.pl, respectively.

[8] There was, however, minor variation across badges and countries (please see Tables 4, 5, and 6 in the appendix).

Nonetheless, it is possible that the average masks significant between-country variability. A multi-level model revealed an intraclass correlation of just 0.001, which means that the proportion of variance in fake news scores that is accounted for by country is close to zero. In other words, most variability lies within groups (Gelman & Hill, 2007). Accordingly, the likelihood-ratio test indicated a multi-level model was not preferred over a standard linear model ($\chi 2 = 2.59$, $p = 0.05$). A standard Ordinary Least Squares (OLS) model with country as a dummy variable shows that compared to Germany, the average inoculation effect is somewhat (but not much) lower in Greece ($\beta = -0.083$, $p = 0.01$, [95%CI -0.01, -0.15]) and Poland ($\beta = -0.09$, $p = 0.01$, [95%CI -0.02, -0.16]). Figure 3 presents the raw data by country and reveals consistent effect-sizes across cultures with a slightly bigger effect for Germany ($M_{pre} = 2.59$, $M_{post} = 2.13$, $M_{diff} = -0.45$, [95% CI -0.41, -0.50], $d = 0.41$) as compared to Greece ($M_{pre} = 2.36$, $M_{post} = 1.99$, $M_{diff} = -0.37$, [95% CI -0.32, -0.42], $d = 0.36$), and Poland ($M_{pre} = 2.52$, $M_{post} = 2.14$, $M_{diff} = -0.37$, [95% CI -0.31, -0.43], $d = 0.33$).

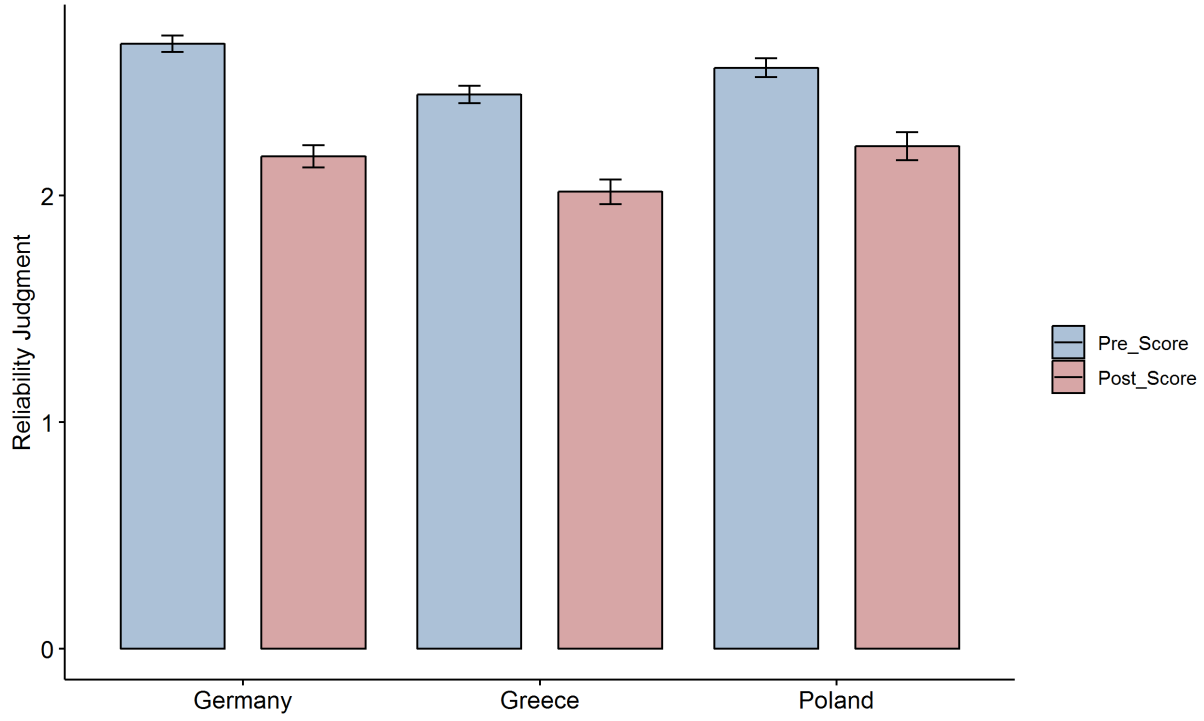

Figure 3. *Average perceived reliability judgments for the fake news scale across countries pre and post gameplay (Germany, Greece, and Poland). Error bars represent 95% confidence intervals. The figure shows that the average perceived reliability of fake news items goes down significantly after gameplay in each country. All pre and post means for the individual items can be found in the methodology appendix.*

Finding 3: The game's effectiveness may not substantially vary across key demographic variables.

The importance of education level and cognitive abilities has been noted as crucial to navigating online information in constructive ways (Nygren, 2019; Nygren & Guath, 2019; Pennycook & Rand, 2019). Previous research has also acknowledged a problematic digital divide between groups in society where the productive use of online information is associated with socio-economic status (Scheerder et al., 2017). Accordingly, we conducted a robustness check by controlling for several demographic covariates for the German, Greek, and Polish versions of the game (the sample size for the pilot in Sweden was too small to control for these variables). Adjusting the model for socio-demographics revealed no main effect for either gender or education but a small significant effect of political ideology and age (see Table 7 in the

appendix). Adjusting for covariates slightly increased between-country differences for Greece (β = -0.09, *p* = 0.01, 95%CI -0.02, -0.16) and Poland (β = -0.12, *p* = 0.01, 95%CI -0.05, -0.20, Table 7, Model 2).

We observe some indications in the pre-test that liberals and people with higher education were a bit more sceptical towards misinformation to begin with. The Polish sample shows the most variation across age, political ideology, and education (see Figures 5-8 in the appendix). For example, conservatives performed slightly better than liberals (meaning: the pre-post-test difference is slightly larger), although this difference is very small and likely explained by conservatives' higher pre-test scores enabling greater learning effects (see appendix for details). On average, we still see that participants with different political ideologies display a significant inoculation effect, highlighting how the intervention may not be hindered by confirmation bias. An exploratory analysis found no significant interaction effects between the intervention, country, and demographics, but since a-theoretical interactions are often underpowered and sample-dependent (Hainmueller et al., 2019) we refrain from further reporting on these here.

These results indicate that the game may work as a broad-spectrum vaccine, as participants with different levels of education, gender, age, and ideologies all learned to better spot misinformation. Nonetheless, it is important to point out that our sample is self-selected and not representative of the respective populations within each country. It is therefore possible that the lack of variation across demographics observed here does not accurately describe the general population. For example, because more liberals opted into the study than conservatives in most countries, it remains unclear to what extent differences in political ideology can be generalized.

We do, however, find a small consistent decrease with age, which is in line with Guess et al. (2019), who found that older people may be more susceptible to online misinformation. Another explanation is that the *Bad News* game was designed with an audience in mind between 15 and 35 years of age, indicating that older people may not have experienced the same level of cognitive involvement as younger people. A third explanation could be rooted in differences in internet literacy, which our survey did not measure. Considering that older internet users are a diverse group with high variance in internet literacy (Hunsaker & Hargittai, 2018), it is possible that the game is effective among older users with high internet literacy and less so for older users who are less familiar with the internet. Our present analysis is unable to disentangle these potential differences.

Methods

The "Bad News" game

In the game, players enter a simulated social media environment and go through 6 scenarios, each representing a different misinformation technique: impersonating people or organizations online; using emotional language to evoke fear or anger; using divisive language to drive groups in society apart (polarization); spreading conspiracy theories; discrediting opponents by using gaslighting and red herrings; and baiting people into responding in an exaggerated manner (trolling). After creating a fictitious "news" site, over the course of about 15 minutes, players must gain followers and build credibility for their site by correctly identifying and making use of these techniques. If their credibility drops to 0 (e.g., by lying too blatantly or using misinformation techniques incorrectly), the game ends. Figure 1 above shows an example of what the gameplay looks like. Although the examples in the game are modeled after real-world events (e.g. impersonating celebrities or politicians online), the game exposes people to severely weakened doses of these techniques in a controlled learning environment (often using humor and purposeful exaggeration, see Figure 1). Consistent with the vaccination metaphor, this process is meant to trigger the "immune system" but not overwhelm it.

The present study: Bad News Intervention

Previous work has shown that inoculation theory can be used to combat various types of online misinformation, for example about conspiracy theories (Banas & Miller, 2013), climate change (van der Linden et al., 2017), and immigration (Roozenbeek & van der Linden, 2018). Crucially, recent work indicates that online games that rely on inoculation theory can be effective at conferring psychological resistance against misinformation *strategies* rather than just individual examples of misleading information (Roozenbeek & van der Linden, 2019). We have dubbed this approach *prebunking*. The most well-known example is the online choice-based browser game *Bad News* (CNN, 2019). The game works by preemptively exposing players to weakened doses of misinformation techniques and combining elements of perspective-taking (stepping into the shoes of someone who is trying to deceive you) and active experiential learning (creating your own media content). By "weakened dose", we mean weakened versions of manipulation techniques in a controlled environment. In other words, "weakened" means strong enough to get people to pay attention (i.e. activate the psychological immune system) but not so convincing as to actually dupe them. We achieved this by 1) using fictional examples throughout the game and 2) by using a combination of humor (Compton, 2018) and extreme exaggeration so that the basic point is still preserved but the risk of duping people is minimized. In the game, players earn 6 "badges", each representing a common misinformation technique (see table 2 in the appendix for all items): impersonating people or groups online (BBC News, 2018b; Goga et al., 2015), using emotional language (Bakir & McStay, 2017; Konijn, 2013), polarizing audiences (Bessi et al., 2016; Melki & Pickering, 2014), spreading conspiracy theories (Lewandowsky et al., 2013; van der Linden, 2013), discrediting opponents and deflecting criticism (A'Beckett, 2013), and online trolling (Griffiths, 2014; McCosker, 2014).

Figure 1. *Example of a fake Twitter post (impersonation technique)*

As part of a collaboration with the United Kingdom's Foreign and Commonwealth Office and the Dutch anti-misinformation platform DROG (2018), we translated *Bad News* into various languages with the goal of building media literacy efforts internationally[9]. In addition, the game was also translated to Swedish in a collaboration between Uppsala and Cambridge University. Following the approach laid out by Roozenbeek & van der Linden (2019), we implemented a voluntary in-game pre-post survey to assess

[9] The game can now also be played in Bosnian (www.getbadnewsbosnia.com), Czech (www.getbadnews.cz), Dutch (www.slechtnieuws.nl), Esperanto (www.badnewsesperanto.com), Moldovan-Romanian (www.getbadnewsmoldova.com), Romanian (www.getbadnews.ro), Serbian (www.getbadnews.rs), and Slovenian (www.getbadnews.si).

players' ability to spot misinformation techniques in each country. We showed participants 6 fake Twitter posts (not featured in the game), each matching one of the techniques mentioned above[10], both before and after playing *Bad News*. Participants rated the reliability of each post on a 7-point scale. These Twitter posts were designed to be realistic, but not real, for two reasons: 1) to exclude familiarity confounds and simple memory-retrieval of "real" fake news, and 2) to have full experimental control over the items' content. This was important because we wanted to make sure the focus of the survey items was on the relevant technique, not the content. Accordingly, we designed each tweet to embed one specific technique (e.g. impersonation) and not another. Figure 4 shows an example of what this looked like in the game environment. We measured the main effects for each language, and conducted paired *t*-tests and regression analyses between average pre- and post- reliability scores for each Twitter post. We also measured socio-demographic variables including age, gender, education level, and political affiliation.

Data collection and sample

Data was collected online through voluntary in-game surveys for a period of 8 months (March-October 2019) for the German, Greek, and Polish versions, and for a period of 13 months (September 2018-October 2019) for the Swedish version (the Swedish game was launched earlier as a pilot study). Participants were recruited organically by driving traffic through media reports linking to the game, as well as through promotional activities conducted by the Foreign and Commonwealth Office, Uppsala University, and local collaborating media literacy organizations (see the "funding" section for more details). The sample is therefore voluntary and self-selected, and consists of people who played the *Bad News* game in Swedish, German, Greek, or Polish and consented to participate in a within-subject (pre-post) in-game survey. As per our ethics application, all incomplete responses, duplicates, and participants under 18 were excluded from the analyses. Table 1 in the appendix gives a detailed description of the sample characteristics for each country. Because we rely on a convenience sampling methodology, participants were more likely to be male, liberal, younger, and educated. However, randomized trials with the English version of the game have been conducted (Basol et al., 2020), highlighting similar results across diverse groups (Basol et al., 2020), consistent with the current findings. Our decision to conduct surveys "in the wild" was based on the real-world nature of the challenge of misinformation. It was therefore important that the game and its translations were freely available online to the entire population in each target country. Lastly, the intervention is of course, not without limitations. For example, it is possible that the intervention has unintended side effects, such as inadvertently promoting fake news. As Tandoc Jr. et al. (2018) note, "people spread fake news with two primary motivations; ideological and financial". We were careful to not provide either incentive in the game (the game itself is politically neutral and monetary incentives are never mentioned). We therefore deem it unlikely that the game will inspire people to produce fake news, but we cannot fully rule out potential side effects, including decay of the inoculation-effect over time. We encourage future research to further explore these issues.

[10] The survey items are listed in the appendix (Table 2).

Bibliography

A'Beckett, L. (2013). Strategies to discredit opponents: Russian representations of events in countries of the former Soviet Union. *Psychology of Language and Communication*, *17*(2), 133-156. https://doi.org/10.2478/plc-2013-0009

Bakir, V., & McStay, A. (2017). Fake News and The Economy of Emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 1–22. https://doi.org/10.1080/21670811.2017.1345645

Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, *39*(2), 184–207.

Banas, J. A., & Rains, S. A. (2010). A Meta-Analysis of Research on Inoculation Theory. *Communication Monographs*, *77*(3), 281–311. https://doi.org/10.1080/03637751003758193

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about Bad News: Gamified inoculation boosts confidence and cultivates cognitive immunity against fake news. *Journal of Cognition*, *3*(1), 1–9.

BBC News. (2018a, February 22). Game helps players spot "fake news." *Www.bbc.co.uk*. Retrieved from https://www.bbc.co.uk/news/technology-43154667

BBC News. (2018b, August 28). A fake billionaire is fooling people on Twitter. *Www.bbc.co.uk*. Retrieved from https://www.bbc.co.uk/news/world-us-canada-45331781

Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., … Quattrociocchi, W. (2016). Users Polarization on Facebook and Youtube. *PLOS ONE*, *11*(8), 1–24. https://doi.org/10.1371/journal.pone.0159641

Carlsson, U. (2019). *Understanding Media and Information Literacy MIL) in the Digital Age: A Question of Democracy*. Retrieved from https://jmg.gu.se/digitalAssets/1742/1742676_understanding-media-pdf-original.pdf

CNN. (2019, July 4). Researchers have created a "vaccine" for fake news. It's a game. *Edition.cnn.com*. Retrieved from https://edition.cnn.com/2019/07/04/media/fake-news-game-vaccine/index.html

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

Compton, J. (2018). Inoculation against/with Political Humor. In J. C. Baumgartner & A. B. Becker (Eds.), *Political Humor in a Changing media Landscape: A New Generation of Research* (pp. 95–113). London: Lexington Books.

De keersmaecker, J., & Roets, A. (2017). "Fake news": Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence*, *65*, 107–110. https://doi.org/https://doi.org/10.1016/j.intell.2017.10.005

DROG. (2018). A good way to fight bad news. *Www.aboutbadnews.com*. Retrieved from www.aboutbadnews.com

European Union. (2018). *Action Plan against Disinformation*. Retrieved from https://eeas.europa.eu/topics/countering-disinformation/54866/action-plan-against-disinformation_en

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Goga, O., Venkatadri, G., & Gummadi, K. P. (2015). The Doppelgänger Bot Attack: Exploring Identity Impersonation in Online Social Networks. *Proceedings of the 2015 Internet Measurement Conference*, 141–153. https://doi.org/10.1145/2815675.2815699

Griffiths, M. D. (2014). Adolescent trolling in online environments: a brief overview. *Education and Health*, *32*(3), 85–87. Retrieved from http://irep.ntu.ac.uk/id/eprint/25950/

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1). https://doi.org/10.1126/sciadv.aau4586

Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice. *Political Analysis*, *27*(2), 163–192. https://doi.org/DOI: 10.1017/pan.2018.46

Human Rights Watch. (2018). Germany: Flawed Social Media Law. Retrieved July 31, 2019, from www.hrw.org website: https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

Hunsaker, A., & Hargittai, E. (2018). A review of Internet use among older adults. *New Media & Society*, *20*(10), 3937–3954. https://doi.org/10.1177/1461444818787348

Klossa, G. (2019). *Towards European Media Sovereignty: An Industrial Media Strategy to Leverage Data, Algorithms, and Artificial Intelligence*. Retrieved from https://ec.europa.eu/commission/sites/beta-political/files/gk_report_final.pdf

Konijn, E. A. (2013). The Role of Emotion in Media Use and Effects. In K. E. Dill (Ed.), *The Oxford Handbook of Media Psychology* (pp. 186–211). Oxford: Oxford University Press.

Kucharski, A. (2016). Post-truth: Study epidemiology of fake news. *Nature*, *540*(7634), 525. Retrieved from http://dx.doi.org/10.1038/540525a

Lee, N. M. (2018). Fake news, phishing, and fraud: a call for research on digital media literacy education beyond the classroom. *Communication Education*, *67*(4), 460–466. https://doi.org/10.1080/03634523.2018.1503313

Lewandowsky, S., Ecker, U. K. H., & 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/https://doi.org/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Gignac, G. E., & Oberauer, K. (2013). The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. *PLOS ONE*, *8*(10), 1–11. https://doi.org/10.1371/journal.pone.0075637

Livingstone, S. (2018, August 5). Media literacy - everyone's favourite solution to the problems of regulation. *Media Policy Project Blog*. Retrieved from http://blogs.lse.ac.uk/mediapolicyproject/2018/05/08/media-literacy-everyones-favourite-solution-to-the-problems-of-regulation/

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*, 101964. https://doi.org/https://doi.org/10.1016/j.gloenvcha.2019.101964

McCosker, A. (2014). Trolling as provocation: YouTube's agonistic publics. *Convergence*, *20*(2), 201–217. https://doi.org/10.1177/1354856513501413

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

McGuire, W. J. (1961). The Effectiveness of Supportive and Refutational Defenses in Immunizing and Restoring Beliefs Against Persuasion. *Sociometry*, *24*(2), 184-197. https://doi.org/10.2307/2786067

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

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

Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake News Detection on Social Media using Geometric Deep Learning. *CoRR*, *abs/1902.0*. Retrieved from

http://arxiv.org/abs/1902.06673

Nygren, T. (2019). *Fakta, fejk och fiktion: Källkritik, ämnesdidaktik och digital kompetens*. Stockholm: Natur & Kultur.

Nygren, T., Brounéus, F., & Svensson, G. (2019). Diversity and credibility in young people's news feeds: A foundation for teaching and learning citizenship in a digital era. *Journal of Social Science Education*, *18*(2), 87–109. Retrieved from https://www.jsse.org/index.php/jsse/article/view/917/1539

Nygren, T., & Guath, M. (2019). Swedish teenagers' difficulties and abilities to determine digital news credibility. *Nordicom Review*, *40*(1), 23–42.

Nyhan, B., & Reifler, J. (2012). *Misinformation and Fact-checking: Research Findings from Social Science*. Retrieved from https://www.dartmouth.edu/~nyhan/Misinformation_and_Fact-checking.pdf

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Reuters. (2019, June 24). Online game helps fight the spread of fake news: study. *Uk.reuters.com*. Retrieved from https://uk.reuters.com/article/us-media-fakenews-game/online-game-helps-fight-the-spread-of-fake-news-study-idUKKCN1TP2X3

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

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

Scheerder, A., van Deursen, A., & van Dijk, J. (2017). Determinants of Internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. *Telematics and Informatics*, *34*(8), 1607–1624. https://doi.org/https://doi.org/10.1016/j.tele.2017.07.007

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News." *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

UNITAR. (2019). Unitar and Partners' Event: Behavioral Insights and Sports - Preventing Violent Extremism. Retrieved January 7, 2020, from www.unitar.org website: https://unitar.org/about/news-stories/news/unitar-and-partners-event-behavioral-insights-and-sports-preventing-violent-extremism

van der Linden, S. (2013). Why people believe in conspiracy theories (What a Hoax). *Scientific American Mind*, *24*, 41–43.

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the Public against Misinformation about Climate Change. *Global Challenges*, *1*(2), 1600008. https://doi.org/10.1002/gch2.201600008

van der Linden, S., & Roozenbeek, J. (2020). A psychological vaccine against misinformation. In R. Greifenader, M. Jaffé, E. Newman, & N. Schwarz (Eds.), *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*. London: Psychology Press.

van der Linden, S., Roozenbeek, J., Oosterwoud, R., Compton, J. A., & Lewandowsky, S. (2018). The Science of Prebunking: Inoculating the Public Against Fake News. In *Written evidence submitted to the Parliamentary Inquiry on Fake News*. Retrieved from http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/fake-news/written/79482.html

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

Wardle, C., & Derakshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Retrieved from https://rm.coe.int/information-disorder-toward-an-

interdisciplinary-framework-for-researc/168076277c
World Economic Forum. (2018). Digital Wildfires. Retrieved August 27, 2019, from reports.weforum.org website: http://reports.weforum.org/global-risks-2018/digital-wildfires/

Funding
The *Bad News* translation project was funded by the United Kingdom Foreign and Commonwealth Office and the University of Uppsala. The authors would like to thank the organizations and individuals involved in this project: DROG, Marije Arentze and Ruurd Oosterwoud in the Netherlands, Wissenschaft im Dialog and Arwen Cross in Germany, MamPrawoWiedzieć and Ewa Modrzejewska in Poland, Ellinika Hoaxes and Andronikos Koutroumpelis in Greece, the Centre for Independent Journalism and Cristina Lupu in Romania, Sorin Ioniță in Romania and Moldova, European Values Think Tank and Veronika Špalková in the Czech Republic, Zavod Časoris and Sonja Merljak Zdovc in Slovenia, the SHARE Foundation and the Centre for Media Law and Kristina Cendic in Bosnia-Hercegovina and Serbia, and Ellen Franzén and Rise in Sweden.

Competing interests
The authors do not report any competing interests.

Ethics
This study was approved by the Cambridge Psychology Research Ethics Committee (registration numbers PRE.2018.085 and PRE.2019.103). Participants provided explicit consent when participating in the study. Three gender identity options were provided in the survey (female, male and other). This study did not hypothesize that there are significant between-gender differences in terms of the size of the inoculation effect, but did control for gender to check this assumption. No significant differences are reported.

Data Availability
All materials needed to replicate this study are available via the Harvard Dataverse at:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MXPKUU

APPENDIX

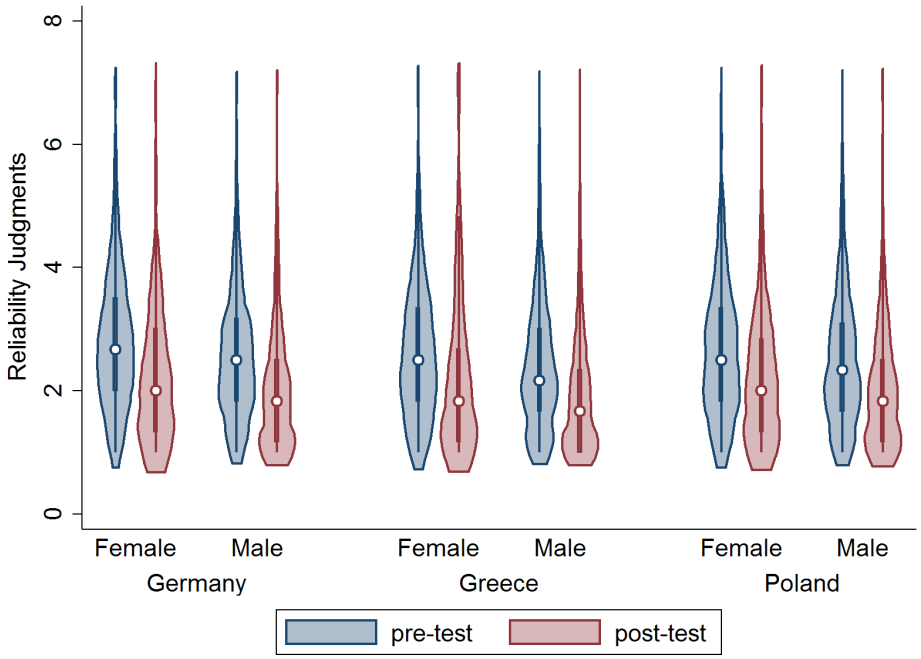

Figure 2. *Violin plots for the German, Greek, and Polish versions of the game (pre-and-post-test), by gender.* The violin shape represents the distribution of the data and the point estimate is the median (with boxplot). All pre-post differences within each country (by gender) are statistically significant at the $p < 0.001$ level.

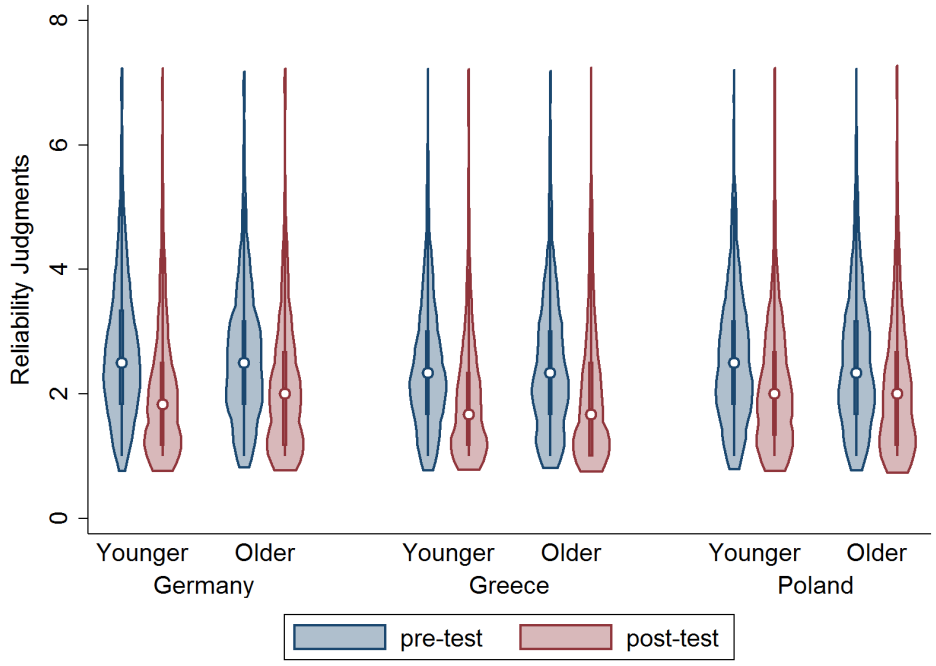

Figure 3. *Violin plots for the German, Greek, and Polish versions of the game (pre-and-post-test), by age (Younger = 18-29, Older = 30-49 & over 50 combined).* The violin shape represents the distribution of the data and the point estimate is the median (with boxplot). All pre-post differences within each country (by age) are statistically significant at the $p < 0.01$ level with the exception for older people in Poland ($p = 0.13$).

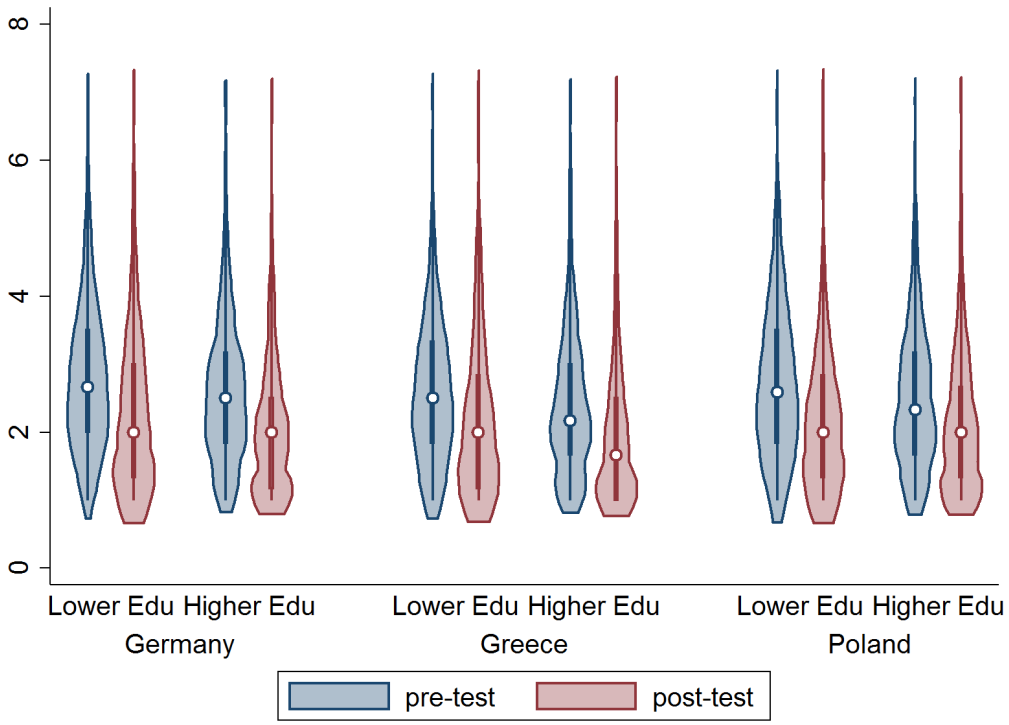

Figure 4. *Violin plots for the German, Greek, and Polish versions of the game (pre-and-post-test), by education (lower education = high school or less).* The violin shape represents the distribution of the data and the point estimate is the median (with boxplot). All pre-post differences within each country (by education) are statistically significant at the $p < 0.001$ level.

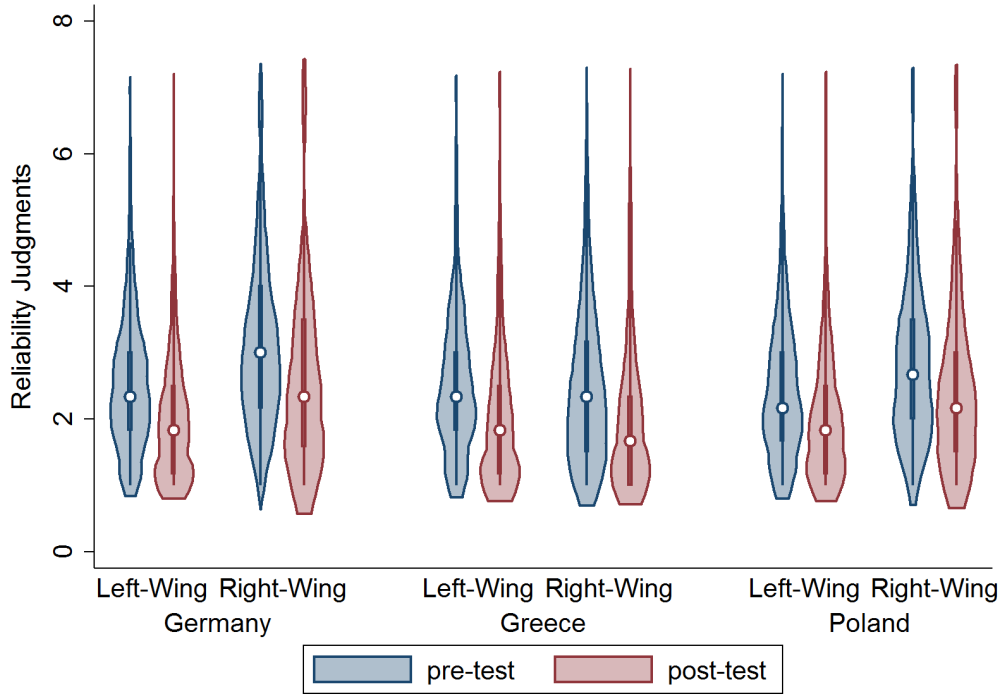

Figure 5. *Violin plots for the German, Greek, and Polish versions of the game (pre-and-post-test), by political ideology (combines "very" and "somewhat" categories).* The violin shape represents the distribution of the data and the point estimate is the median (with boxplot). All pre-post differences within each country (by ideology) are statistically significant at the $p < 0.001$ level.

Table 1. *Sample characteristics*

Demographics	*Sweden* *n*	*Germany* *n*	*Greece* *n*	*Poland* *n*
Gender	428	4,134	3,225	2,942
Male	0.45	0.63	0.75	0.58
Female	0.46	0.25	0.20	0.35
Other	0.09	0.12	0.05	0.07
Age	428	4,101	3,213	2,929
18-29	0.36	0.32	0.31	0.46
30-49	0.46	0.52	0.56	0.45
Over 50	0.18	0.16	0.13	0.09
Education	425	4,019	3,199	2,917
High school or less	0.20	0.17	0.21	0.15
Higher degree	0.14	0.24	0.37	0.22
Some college/university	0.66	0.59	0.42	0.63
Political Ideology	NA	4,066	3,182	2,910
Very left-wing		0.05	0.09	0.04
Left-wing		0.28	0.16	0.18
Somewhat left-wing		0.34	0.25	0.30
Neutral		0.21	0.31	0.32
Somewhat right-wing		0.07	0.14	0.11
Right-wing		0.02	0.03	0.03
Very right-wing		0.03	0.02	0.02

Note**:** Sample characteristics (proportions) for all respondents who filled out the sociodemographic questions at the end of the in-game survey. Note that because the survey is voluntary, sample sizes differ by item, country, and demographics.

Table 2. *Survey items (fictitious Twitter posts); participants were asked to assess the reliability of each Twitter post on a 1-7 Likert scale, both before and after playing.*

Item name	English	Swedish	German	Greek	Polish
Impersonation	The 8th season of #GameOfThrones will be postponed due to a salary dispute	Vi ber om ursäkt för #GameOfThrones säsong 8 och sänder i höst en ny avslutande säsong 9.[11]	Wegen Lohnverhandlungen wird die 8. Staffel von #GameOfThrones verspätet ausgestrahlt.	Η 8η σεζόν του #GameOfThrones θα αναβληθεί λόγω μιας διαφωνίας αναφορικά με τους μισθούς.	Kolejny sezon serialu #GraOTron zawieszony, w tle – spór o pieniądze.
Emotion	NEWS ALERT: Baby formula linked to horrific outbreak of new, terrifying disease among helpless infants. Parents despair	-	EILMELDUNG: giftiges Essen verursacht FURCHTBAREN Ausbruch einer Krankheit unter hilflosen Kleinkindern. Eltern verzweifelt.	ΕΚΤΑΚΤΟ: Βρεφική φόρμουλα προκάλεσε ΦΡΙΚΤΗ έξαρση καινούργιας τρομακτικής ασθένειας σε αβοήθητα νεογέννητα. Οι γονείς απελπίζονται.	PILNE: Mleko modyfikowane wywołuje nieznaną dotąd, STRASZNĄ chorobę u niemowląt. Rodzice są przerażeni!
Polarization	The myth of equal IQ between left-wing and right-wing people exposed #TruthMatters	-	Mythos der gleichen Intelligenz von links- und rechtsgesinnten Menschen aufgedeckt #WahrheitZählt	Ο μύθος του «ίσου IQ» μεταξύ αριστερών και δεξιών εκτίθεται. #Η_αλήθεια_μετράει	Obalono mit takiego samego IQ u ludzi o lewicowych i prawicowych poglądach. #PrawdaSięObroni
Conspiracy	The Bitcoin exchange rate is being manipulated by a small group of rich bankers #InvestigateNow	Växelkursen för Bitcoin blir manipulerad av en liten grupp rika bankirer #UndersökNu	Reiche Investmentbanker manipulieren den Wechselkurs des Bitcoin #UntersuchungJetzt	Η αξία του Bitcoin παραποιείται από μικρή ομάδα πάμπλουτων τραπεζιτών. #ΔιερεύνησηΤώρα	Kursy bitcoina są zmanipulowane przez wąską grupę bogatych bankierów. #ŚledztwoTrwa
Discredit	The mainstream media has been caught in so many lies that it can't be trusted as a reliable news	PK-median har blivit avslöjade med att ljuga så många gånger så det går inte att se dem	Die Mainstream-Medien wurden so vieler Lügen überführt, dass sie keine vertrauenswürdige Quelle mehr sind. #FakeNews	Τα συστημικά ΜΜΕ έχουν πιαστεί στα πράσα τόσες φορές που δεν μπορούμε να τα εμπιστευόμαστε πια. #ΨευδείςΕιδήσεις	Mainstreamowe media złapano na tak wielu kłamstwach, że nie można ich traktować jako wiarygodne źródła wiadomości. #FakeNews

[11] This item was adapted slightly in May 2019 for the Swedish version. The original item read "The 8th season of #GameOfThrones will be postponed due to a salary dispute".

	source #FakeNews	som en trovärdig nyhetskälla #FakeNews			
Trolling	Another shark loan for developing countries @WorldBank #WorldOfExtortion #HumanBanking	-	Noch ein unfaires Darlehen in den globalen Süden @WorldBank? #BankenMenschlichMachen	Κι άλλος τοκογλύφος για τις αναπτυσσόμενες χώρες @WorldBank; #ΠαγκόσμιαΤράπεζαΕκβιασμού	Kolejna lichwiarska pożyczka dla krajów rozwijających się @BankŚwiatowy? #BankŚwiatowegoZdzierstwa #LudzkiBank

Table 3. *Average reliability (pre-post) judgments overall and for each fake news badge by country (Sweden).*

	M_{pre}	M_{post}	M_{diff}	95%CI_{diff}	Cohen's *d*	*n*
Fake news scale	2.44	2.12	-0.32	[-0.13, -0.51]	0.24	173
Impersonation	3.10	2.68	-0.43	[-0.07, -0.78]	0.18	174
Conspiracy	2.41	2.10	-0.31	[-0.12, -0.50]	0.17	379
Discrediting	2.07	1.79	-0.27	[-0.08, -0.47]	0.15	376
Trump $_{(control1)}$	6.29	6.32	0.03	[-0.21, 0.15]	0.01	380
Brexit $_{(control2)}$	5.99	6.10	0.11	[-0.36, 0.14]	0.07	134
Literacy $_{(control3)}$	5.20	5.15	-0.05	[-0.32, 0.22]	0.02	244

Table 4. *Average reliability (pre-post) judgments overall and for each fake news badge by country (Germany).*

	M_{pre}	M_{post}	M_{diff}	95%CI_{diff}	Cohen's *d*	*n*
Fake news scale	2.59	2.13	-0.46	[-0.41, -0.50]	0.41	2,038
Impersonation	3.17	2.61	-0.56	[-0.47, -0.65]	0.26	2,076
Polarization	2.24	1.94	-0.30	[-0.22, -0.37]	0.17	2,065
Conspiracy	3.18	2.44	-0.74	[-0.66, -0.83]	0.38	2,073
Emotion	1.75	1.50	-0.24	[-0.18, -0.30]	0.18	2,068
Discrediting	2.10	1.84	-0.25	[-0.18, -0.32]	0.15	2,068
Trolling	3.09	2.44	-0.65	[-0.56, -0.73]	0.34	2,063

Table 5. *Average reliability (pre-post) judgments overall and for each fake news badge by country (Greece).*

	M_{pre}	M_{post}	M_{diff}	95%CI_{diff}	Cohen's *d*	*n*
Fake news scale	2.36	1.99	-0.37	[-0.33, -0.42]	0.36	1,518
Impersonation	2.80	2.20	-0.60	[-0.49, -0.72]	0.29	1,539
Polarization	1.88	1.70	-0.18	[-0.10, -0.26]	0.12	1,539
Conspiracy	2.26	1.96	-0.30	[-0.22, -0.38]	0.18	1,534
Emotion	1.31	1.29	-0.02	[-0.04, 0.08]	0.02	1,535
Discrediting	3.43	2.67	-0.76	[-0.65, -0.86]	0.34	1,534
Trolling	2.49	2.11	-0.38	[-0.47, -0.29]	0.21	1,534

Table 6. *Average reliability (pre-post) judgments overall and for each fake news badge by country (Poland).*

	M_{pre}	M_{post}	M_{diff}	95%CI_{diff}	Cohen's *d*	*n*
Fake news scale	2.52	2.14	-0.37	[-0.32, -0.42]	0.33	1,332
Impersonation	2.67	2.26	-0.40	[-0.30, -0.51]	0.20	1,355
Polarization	2.24	1.96	-0.28	[-0.18, -0.38]	0.16	1,350
Conspiracy	2.75	2.30	-0.44	[-0.34, -0.54]	0.24	1,352
Emotion	1.51	1.42	-0.09	[-0.02, -0.16]	0.08	1,355
Discrediting	3.31	2.75	-0.57	[-0.46, -0.67]	0.28	1,349
Trolling	2.61	2.18	-0.43	[-0.33, -0.53]	0.25	1,348

Table 7. *Average fake news reliability (pre-post) judgments OLS regression with and without covariates.*

Post-Pre (Diff)	Fake News Reliability No Covariates (β)	Fake News Reliability With Covariates (β)
Country (ref: Germany)		
Poland	-0.083** (-0.02, -0.15)	-0.095*** (-0.03, -0.16)
Greece	-0.086** (-0.02, -0.16)	-0.128*** (-0.06, -0.20)
Gender (Male)	-	-0.047 (-0.11, 0.02)
Age (Older)	-	-0.082*** (-0.03, -0.13)
Education (Higher)	-	0.011 (-0.01, 0.03)
Ideology (Right-Wing)		0.026* (-0.00, 0.05)
N	4,888	4,543
Adj. R2	0.001	0.006
Δ adj.		0.005
$F_{(change)}$	4.27	5.43

Note: * p <0.05, ** p = 0.01, ***p < 0.001. OLS regression estimates by country (Model 1) without covariates and with covariates (Model 2). 95% confidence intervals are provided in parentheses. In Model 1, the beta-coefficients (unstandardized) simply represent the mean difference with Germany (i.e. 0.08 points lower). For variable coding please see Table 1.