ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

AUTHOR RESPONSE

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Nature Humanities and Social Sciences Communications* **5**, 65.

*response by*

**Sander van der Linden**, University of Cambridge

**Jon Roozenbeek**, King’s College London

*Review version 1.0 (DATE)*

*Review template version 1.0*

*License: CC BY 4.0*

We greatly appreciate the time invested by Steve Haroz (Reviewer) and Ian Hussey (Recommender) in evaluating our manuscript for potential errors in such a detailed and thoughtful manner. We are big fans of the ERROR project and volunteered immediately when asked to include our paper and are pleased to respond here to the conclusions of the report about our 2019 paper. In general, our reading is that the report identified a few minor valid (reporting) errors that are worth noting but don’t otherwise alter the conclusions or results from our paper. We are pleased the reviewer’s additional analyses reinforced or supported our initial conclusions. The main issue seems to revolve around potential order effects of items in the game. We discuss this and all other issues below.

**Design**

1. Attention check: The reviewer notes that we did not use attention checks, and the recommender commented that this remark can be ignored because it is not an error. We agree on the usefulness of attention checks and are pleased to say that we have indeed implemented these in our later studies, including a passcode for verification of game completion (Leder et al., 2024; Maertens et al., 2021).
2. Order effects (data generation): The reviewer spent most of their time and error report on this issue so it is worth covering this point in detail. As noted to the reviewer, the order in which participants saw both the “real news” (control) and “fake news” items in the game was fixed not randomized with the control items interspersed between the fake items. The reviewer presents an interesting simulation which offers an overview of the expected pattern in the data if the ordering of the items were fixed, namely that if participants answered say item 2, it must be the case that there is also data for the preceding item 1 given that participants could not skip questions. In other words, it would not be possible that data is recorded for item 2 but not 1. The simulation showed that this is indeed the case for certain ranges of the data but not for others where it indeed appears as if those who answered item 2 don’t have data for item 1 or those who answered item 3 don’t have data for item 2 and so on. The reviewer is correct in this assessment (the simulation versus actual data comparison makes sense).

We checked with the software developers on this oddity and they clarified what the issue was. First, it is important to mention that in the game’s content management system (CMS) the presentation of the item order was always fixed as reported and that’s how all participants *viewed* the items. However, when the game launched, it went viral (e.g., see [Reddit](https://www.reddit.com/r/science/comments/c5ptfz/fake_news_vaccine_works_suggests_a_large_new/), 2019) and the backend was overwhelmed with data storage requests so that not all incoming data was recorded (e.g. due to time lags). It therefore appears that responses to items were sometimes not recorded in the backend of the game due to data storage glitches, which makes it so that even though a participant filled out all items in a fixed order, data on some of the items were not stored appropriately. This explains why for some participants, a response was recorded for (for example) item 2 but not item 1, and so on. The developers have assured us that it would not be possible for the order to have been different given that the CMS doesn’t have order randomization capabilities. The launch was exploratory at the time and we were not prepared for the huge data storage requests when the game went viral so unfortunately not all possible data were recorded. However, the paper only reports on pre-post completes for a given technique and the supplement already provided missing-at-random tests and predictors of missingness for the key variables (see Supplement, Tables S1-S2).

1. Order effects (design): Perhaps a more pertinent question is why the order in which the items were presented matters in the first place. The main point is that the order wasn’t randomized which in within-subject designs could potentially raise the concern of order, testing, or carry-over effects. We think these concerns are unlikely to pose a problem, for reasons we will spell out in detail. However, we take the point that we did not state the order in which the items were presented in the published paper. The reviewer points to the JSON file of the current version of the game which contains the item ordering. We asked the developers (we were not aware of the existence of this file) and they confirmed the validity of this file but they don’t have the JSON file from the original Bad News version (+6 years ago) anymore. We did, however, screenshot the items themselves for the purpose of preservation which we include in Appendix A exactly as they were shown in the game (which also addresses the reviewers’ concern about the exact stimuli, though we already reported the item wording verbatim in the paper). Although we recognize the benefits of item randomization, this wasn’t a technical capability for us due to the limitations of the CMS. At the time, we took inspiration from statistician Andrew Gelman’s posts, which suggested psychologists should focus more on large sample within-subject designs for precise estimates of within-person change without worrying about “poisoning the well”–statistically this being preferable to the downsides of noisy (often small N) between-subject designs (Gelman, 2017). We don’t see this design choice as an error but more as part of on-going debates about methodological tradeoffs, including how to best organize the items. Moreover, we note that any choice we would have made at the time (whether to keep the pre-post item order the same or switch it up) had both upsides and downsides. What *is* an error, of course, is our failure to mention the item ordering.

We wish to emphasize here that we did immediately explore item effects, testing effects, and within subject versus between subject differences in the Bad News game through a series of follow-up publications (Roozenbeek et al., 2021). In one paper, we specifically adopted a Solomon three-group design to see if there’s a difference between the pre-post design and a post-only control group, which ruled out learning and testing effects (Roozenbeek et al., 2021). We have also run studies where we randomly administered different item sets pre- and post-gameplay, which did indicate some small items effects which relate to the plausibility/difficulty of the items but no order effects (Roozenbeek et al., 2021, 2022; Maertens et al., 2021). In addition, we ran studies where we used fictional items (which we created ourselves) versus social media items taken from real-world misinformation (Roozenbeek and van der Linden, 2021; Roozenbeek, Traberg et al., 2022) and randomized versus non-randomized orders (Roozenbeek et al., 2021). To briefly summarize these studies, we are confident that the order of the items does not matter for the inoculation effect observed in the Bad News game. What does matter is potential *item* effects (but again not their ordering).

1. Ethics (design): The reviewer didn’t think our reasons for not having a control group at the time were very convincing. The recommender notes this is not an error but we nonetheless wish to clarify, as stated in the paper, that we were simply following ethics protocol. The university was concerned that going live with an educational game could generate complaints if it was only available to some people. We are governed by the clinical school, whose reviewers often think about whether it is ethical to deprive a control group of a treatment (for example in the case of a medical treatment) We agree this is not usual for psychology, but these were the conditions of our ethics approval. We did also get ethics approval for subsequent smaller N randomized experiments in a controlled setting (e.g., Maertens et al., 2021; 2025). The reviewer seems to disagree with this reasoning in that the educational benefit of the intervention wasn’t clear a priori (which is true), but again it wasn’t up to us. The reviewer offers an alternative no placebo design but this again relies on a randomization function which the initial game didn’t have and we didn’t have resources at the time to have it developed so that’s why we stuck with the simplest possible option and design. In terms of cross-category effects, we tested this later on by selectively switching off techniques in the game and positive cross-protection effects exist, though they are small (Roozenbeek et al., 2022).
2. Pre-treatment values (design): The reviewer notes that the control or “real news” items have a very different distribution than the fake news items. This is correct and makes sense because the data generation process underlying the production of fake and true news is very different (Vosoughi et al., 2018). In pretty much every study on misinformation you will see that people rate real news as more credible than fake news because people have a baseline ability to discern between the two (Sultan et al., 2024). They are not meant to be similar otherwise it wouldn’t be ecologically valid. This is related to the reviewers’ concern of how baseline differences in the items might impact the intervention, but we are not treating the real and fake news items as part of a factorial design; instead, these were just exploratory items that we look at separately (but more on that under “analysis”). This also helps answer the reviewers’ question about the polarization item not having a “control”. None of the fake items have a counterpart control – the real news items are a separate set. In later research (e.g., Roozenbeek et al., 2022) we did eventually create a balanced item set with (say) 10 manipulative items and their 10 non-manipulative counterpart controls so we could formally look at “discernment” between the two (which is what the reviewer is getting at) but we didn’t do that in this first study, which is a limitation but not an error. Nonetheless the reviewers’ interaction analysis (pre-post \* fake vs real) returns the same result anyway which is great to see and validates the initial conclusions further.

**Results**

1. Analyses (descriptives and main effects): We are pleased the reviewer reproduced the main effects as reported. We indeed clarified that the 95% confidence intervals are based on the standard error, as is usual in psychology. We are happy to clarify this in the paper if desirable.
2. Analyses (effect-sizes): The reviewer notes we reported effect-sizes in the same direction, which is true because we didn’t think the sign mattered, but the reviewer is correct that for control (real news) item 1 the sign should be positive and negative for all others. We are happy to correct the sign for the effect-sizes. The reviewer also notes that the p-value for control item 1 also has greater significance than initially reported (p = 0.0002 vs p = 0.002) – we indeed accidentally dropped a zero when transcribing, thanks for this, the p-value should be 0.0002. We are also happy to correct this. The reviewer is confused as to why we reported BOTH Cohen’s *Dz* and Hedges’ *g* for all effects. The answer is simply transparency, as the reviewer notes they use different pooling methods (but converge for large N) so we wanted to show effect-sizes using both methods so the reader can form their own impression.
3. Analyses (ordinal estimation): The reviewer notes we treated ordinal data as numerical. The recommender noted this is not an error but a matter of methodological debate. We agree with both points. For clarity, the analyses are performed on averaged data (e.g., 3.2, 3.6, 3.7) which approximates numerical data pretty well but we should note that in future work we have also modelled the response options using Bayesian and ordinal probit models to account for the ordinal structure of the data (Leder et al., 2024). We are also reassured to see the reviewer’s paired ordinal regression reproduced the same conclusions as reported.
4. Analyses (MANOVA): The reviewer wonders why we reported a MANOVA. This was simply to control for the family wise error rate before doing the univariate testing across multiple DVs but we agree this was not essential either way.
5. Analyses (interaction): The reviewer points to the well-known article “the difference between “significant” and “not significant” is not itself statistically significant”. We are well aware of this problem but don’t necessarily agree that we interpreted our results this way. The control/real news items items do not match the fake news items, neither in content nor in number, so this is not a factorial design. The two control (real news) items were exploratory and we looked at two different outcomes, a pre-post effect on the controls and a pre-post effect on the fake news items. We state that one effect was significant and the other wasn’t but we don’t make conclusions about the difference. It is perfectly valid to report on two categories separately. If the reviewer thinks we are ultimately interested in the interaction, that’s an interesting question, but not one we initially set out to answer. For example, in the reviewers’ example, they compare the impersonation items vs a control/real news item but this doesn’t make much sense conceptually because they aren’t balanced/paired that way. Of course, it’s reassuring that the reviewer’s interaction approach reproduces the same conclusions, but we just wanted to point out why we didn’t do this in the paper. Another way to look at this problem is through the lens of “discernment” – which is now a common outcome measure in the literature and gets at the same idea where you subtract the pre-post difference for real news items from the pre-post difference of the fake news items (the “difference-in-difference”). If you do that for our data, you get a significant overall improvement in discernment (Mdiff = 0.55, 95%CI; 0.53, 0.58, *d* = 0.37, *p* < 0.001). Unsurprising perhaps as the mean difference for control items is basically 0 pre-post. However, discernment by itself is not transparent as discernment can go up because people improve in their fake news recognition, real news recognition, or some combination of both (where a decrease in one can still lead to higher discernment if the decrease is offset by a higher gain in the other variable). This outcome measure wasn’t used or widely adopted in 2018. We therefore reported real news and fake news recognition separately, from which it was already obvious that there was significant discernment as people improved in their recognition of fake news with no change in real news. Given that the items were not matched we didn’t feel it was appropriate to report on discernment or the interaction but it’s good to know that the results remain the same either way.
6. Analyses (prior susceptibility): The reviewer comments on the reported finding that those with the lowest prior score (highest misinformation “susceptibility”) improved the most. Noting that our median split may have been an oversimplification. The reviewer preferred an ordinal ANOVA with a pre-test interaction, which encouragingly again reproduces the same result. We agree that the ordinal ANOVA produces an interesting visualization that illustrates potential floor and ceiling effects. Of course, it is true that given the ordinal nature of the scale, if people rate the reliability of fake news highest on the pre-test, they can only go down on the post-test and vice versa. However, it is important to keep in mind that people do not *need* to adjust their ratings downward; they can keep their rating the same on the post-test as on the pre-test if they so choose. If the intervention effectively changes how people evaluate fake news then those with the highest susceptibility pre should improve the most because they have the most room to move on the scale – but if the intervention is not as effective or if they don’t learn from it, they don’t have to move or adjust their ratings. So, it is a combination of being constrained by the start and end points of an ordinal scale but also by the fact that people did move in response to the intervention. Large majorities of pre-post pairings on the control/real items did not move (the median, 25th, 50th, and 75th percentile of the data distribution are nearly all the same pre-post). Potential floor and ceiling effects are also clear from the density distributions which we already reported in Figures 6 and 7 of the paper. We weren’t familiar then with the ordinal ANOVA package in R (good suggestion) but a median split on the high and low categories, though perhaps less efficient, is perfectly acceptable and a matter of methodological preference/debate (see Iacobucci et al., 2015 the median split: robust, refined, and revived), so we don’t see this as an error.
7. Analyses (sample sizes): We don’t fully grasp what the reviewer was struggling with here. They claim they cannot reproduce the sample sizes or results for the demographic breakdowns. Any point-and-click software using our dataset will reproduce the exact sample size, degrees of freedom, and means exactly as reported. Here’s an example for Ideology (the first one) on the reviewer’s list, created using Jamovi (jamovi.org).

A screenshot of a test

AI-generated content may be incorrect.

These data match the reported results exactly “conservatives rated fake headlines more reliable than liberals at pre-test, *M*cons = 2.85 vs. *M*lib = 2.38, *M*diff = 0.47, [95% CI 0.52, 0.42], *t*(14032) = 19.66, *p* < 0.001”. The same is true for all other breakdowns. The total sample sizes for the socio-demographic variables also match the total sample sizes listed in the supplement. Perhaps the reviewer got confused because of the lack of guidance on what we did, but the recoded categories/bins are also all provided in the data sheet. We did catch one error that maybe threw the reviewer off: they note the impossible small sample size of “1167” – which is indeed a typo and should have been “1167**4**” – a digit got dropped somehow (see output below).

A screenshot of a test

AI-generated content may be incorrect.

The reviewer does make a good point that in the abstract there appears a coarse approximation of N = 15,000. We had to combine the extra polarization badge data collection with the main study but this rounding may have been a proofing/copyediting error. It should have been 14,266 + 885 (polarization) = 15,151. The reviewer also makes the correct observation that we listed the highest N for completed pre-posts (14, 266) but the reviewer thinks we should include a higher number (14, 536) which represents the number of participants in the dataset for at least 1 item with completed pre-posts. We don’t have a strong view on this. This number doesn’t take the polarization data into account, however. That would be 14,536 + 855 = 15,391. We are happy to report this new number and ask the journal to correct the abstract. Other suggestions are welcome too.

The reviewer also correctly spotted a sample size error for the polarization badge. The article says 885 but this is a typo as the supplement says 855 which is also what the reviewer claims and what the dataset returns. The 1,770 paired refers to the total number of *responses* (2\*885 = 1,770) NOT the total number of people. We think this caused a lot of the confusion for the reviewer. The reviewer is correct that the total sample is 3,980. We think the easiest solution is just to report the correct n = 855 paired responses.

Other concerns: The reviewer raises a number of other concerns about data transparency, code, and the error with activating the polarization technique. First we fully recognize that the first time we ran this study using the newly-built game, a lot was exploratory and we had technical issues. We had the software developers export the data and a statistician process the data in the right format. After we recognized the pipeline issues, we wrote transparent code to transform the data once exported so the raw data, coding scripts, and clean data are all available on the OSF (<https://osf.io/48z2d>). Our game data has been the subject of many replications and meta-analyses because we are committed to open science and data. We also acquired funds to implemented an in-game survey data collection tool that allowed for a dashboard to output the data in the right format to help streamline the whole process. This is just to say that we learned a lot from the initial study and acted swiftly to design a more transparent process by which we and other scholars using the game can export and use the data.

The reviewer questions technical issues surrounding storage and the size of the files. The JSON file the reviewer is looking at doesn’t cover much data: the raw data files are large, there’s hundreds of thousands of people filling out partial survey items (possibly many more during the initial launch), so much so that our servers crashed shortly after the launch. On Reddit, this was referred to as the “Reddit hug of death” (a website crashing due to an excess in web traffic redirected there from the site). We are glad the reviewer mentions they’ve made their fair share of data collection mistakes over the years too, it happens! The reviewer also thinks we could have included the polarization data if we had caught the error sooner. This is 20/20 hindsight of course. What happened was that the polarization item has quotation marks around the text (“) which prevented the game from recording the data because it didn’t recognize the string. It took us a few days to figure this out (we didn’t realize the quotation marks were the issue) and after the game went viral on launch day, participation steeply dropped so by the time the polarization data collection was re-activated we had lost momentum in terms of data collection. You live and learn!

We appreciate the time the reviewer invested in checking our data and results. The tests themselves were so simple (t-tests) that we just used point-and-click software to generate the results and most of the Figures – we didn’t have any code – the data was delivered to us in the right format by the developers/stats consultant and we used JASP/JAMOVI to run most of the analyses (with some graphs and effect-sizes done in STATA). As mentioned, after this study, we coded everything in R and deposited the data cleaning and analysis pipeline on the OSF pages, along with the raw data, code, and clean data. So we understand this critique about the initial paper but have acted accordingly to rectify it.

**Conclusion**: We thank the reviewer and recommender again for their time and thoroughness in identifying several minor issues with the published paper. Whilst we deem many of the reviewers’ comments as differences in opinion about measurement, in particular, we have flagged the following true errors; (1) a p-value was more significant than appeared (p < 0.0001), the sample size in the abstract is an approximation not exact should be (N = 15, 391), the polarization sample size contained a typo (n = 855 NOT 885), another sample size typo 11674 not 1167), direction/sign of effect-sizes are missing, and we did not mention the order of the items in the original manuscript (fixed order not randomized) including an issue with the data generation process (more responses than could be recorded). In the interest of correcting the record and transparency, we are happy to contact the journal to issue a correction to clarify these issues.

**References**

Gelman, A. (November 27, 2017). Poisoning the well with a within-person design? What’s the risk? Statistical Modeling, Causal Inference, and Social Science.

<https://statmodeling.stat.columbia.edu/2017/11/25/poisoning-well-within-person-design-whats-risk/>

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., & Popovich, D. L. (2015). Toward a more nuanced understanding of the statistical properties of a median split. *Journal of Consumer Psychology*, *25*(4), 652-665.

Leder, J., Schellinger, L. V., Maertens, R., van der Linden, S., Chryst, B., & Roozenbeek, J. (2024). Feedback exercises boost discernment of misinformation for gamified inoculation interventions. *Journal of Experimental Psychology: General*, *153*(8), 2068.

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.

Maertens, R., Roozenbeek, J., Simons, J. S., Lewandowsky, S., Maturo, V., Goldberg, B., ... & van der Linden, S. (2025). Psychological booster shots targeting memory increase long-term resistance against misinformation. *Nature Communications*, *16*(1), 2062.

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). Disentangling item and testing effects in inoculation research on online misinformation: Solomon revisited. *Educational and Psychological Measurement*, *81*(2), 340-362.

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), eabo6254.

Roozenbeek, J., & van der Linden, S. (2020). Breaking Harmony Square: A game that “inoculates” against political misinformation. *The Harvard Kennedy School Misinformation Review*.

Roozenbeek, J., Traberg, C. S., & van der Linden, S. (2022). Technique-based inoculation against real-world misinformation. *Royal Society open science*, *9*(5), 211719.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, *359*(6380), 1146-1151.

**Appendix A**

Real (control) Polarization

A screenshot of a message

AI-generated content may be incorrect.A screenshot of a phone

AI-generated content may be incorrect.

Impersonation Conspiracy

A screenshot of a message

AI-generated content may be incorrect.A screenshot of a social media account

AI-generated content may be incorrect.

Discrediting Real (control)

A screenshot of a message

AI-generated content may be incorrect.A screenshot of a social media post

AI-generated content may be incorrect.