

ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

ERROR REPORT

Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5, No. 65.
<https://doi.org/10.1057/s41599-019-0279-9>

reviewed by

Steve Haroz

May 27, 2024

Annotations by the Recommender

Ian Hussey

Jan 28, 2025

For the sections below, indicate whether you discovered any errors using the dropdown menu ▾. Describe the **errors** you discovered, the **methods** that you used to find them, and the **amount of time** you invested in the search. Refer to specific files to allow verification of your review. For the assessment below, make sure to check if authors have provided **supplementary analyses** as these may clear concerns arising from the (interpretation of) the primary analyses. If you have written code yourself for the review, please attach it to the report. Please indicate the version of software/packages you used to run the original code and/or your own code.

Notes from the Reviewer:

- I labeled each concern using its corresponding section number from the complete analysis, which can be downloaded at:
<https://github.com/ianhussey/error-reviews/tree/main/reviews%2C%20responses%2C%20and%20recommendations/van%20der%20Linden%20et%20al%20-%202019%20-%20Fake%20news%20game%20confers%20psychological%20resistance%20against%20online%20misinformation/1%20review> (see [fake-news-reanalysis-edited.html](#)).
- I am reporting concerns using two (somewhat subjective) levels:
 - Concern or error that does not appear to substantially impact the conclusions
 - Concern or error that calls the conclusions into question

I. METHODS, MEASUREMENT, AND DESIGN

1. Design Errors found ▾

Are there errors in the conceptual design of the study? E.g., flawed randomisation technique

■ 4.2 - Same response every time - Without any check that subjects are actually reading and responding to the prompts, inattentive responses such as picking the same number repeatedly may be cause noise or bias in the results.

- Note by Recommender: Attention checks improve the validity of claims, but their omission does not constitute an error.

🔴 9.1 - Order effects - The article does not mention how question order was balanced. According to [recommender clarification: correspondence with] the authors, the order was fixed for all subjects with control and intervention questions interspersed. However, despite substantial effort, I was unable to reproduce the order, which should have been apparent from the pattern of incomplete responses due to dropouts. Instead, the order appears to be inconsistent. It is unclear why the question order varied, but neither fixed ordering nor a per-subject random ordering could explain the results.

🔴 9.2 - Confounded controls - The pre-test control responses are substantially different from the pre-test intervention responses. Consequently, it is unclear if the differences in the change from pre-test to post-test is a result of the intervention or of whatever about the question caused the initial difference.

🔴 9.3 - Unconvincing ethical concerns - The article claims that the common approach of giving subjects a placebo treatment for comparison with the intervention would be unethical. The reasoning is questionable, but it could be worked around with a more balanced study design.

- Note by Recommender: The absence of a control condition undermines the validity of claims, but debate about the ethics of a procedural feature that was not used does not constitute an error.

🔴 10.2 - Polarisation study's lack of control questions - The polarisation study does not report any controls that could be used for comparison.

Time spent: 20 hours (mostly on the strange order effects issue)

2. Measurement Errors found ▾

Are there any measures, techniques, or devices that were incorrectly applied or inappropriate for the specific task described in the paper?

🟡 6.4 - Ordinal, not numerical - The t-tests treat the Likert responses as numerical despite them being ordinal. I reanalyzed the analyses using ordinal regression instead of t-tests. All ordinal p-values were categorically similar to the t-test p-values, so this issue did not appear very consequential.

- Note by Recommender: The choice to model ordinal data as continuous is a matter of statistical debate and preference and does not constitute an error.

See also:

7.5 - Ordinal and comparison

- Note by Recommender: See previous re ordinal modelling. For the note of 'comparison', see section "4. Model Misspecification" for more discussion.

10.1.4 - Ordinal, not numerical

- Note by Recommender: See previous re ordinal modelling.

Time spent: 5 hours

3. Preregistration Consistency Not applicable ▾

Are there substantial deviations from the preregistration, particularly undisclosed ones?

Time spent:

4. Sampling Errors found ▾

Is there an error in the sampling strategy? Is the power analysis reproducible? Does the model used for the power analysis match the model in the substantive analyses? Were separate power analyses conducted for all primary analyses?

I could not reproduce many of the reported sample sizes. Some that I could reproduce were found to have missing exclusion criteria, incorrectly reported exclusion criteria, typos, or even combinations of those problems (see section 5.1).

I was able to reproduce some of the sample size reporting errors by trial-and-error of possible exclusion criteria until the right number was found (which required a lot of time). These exclusion criteria were not necessarily problematic, but not reporting them is very concerning. In other cases, I could not reproduce the reported sample size.

Sections:

5.1 - Response count

5.2 - Paired response count

- 7.2 - N is not reproducible
- 8.1 - Demographic breakdown (all subsections)
- 10.1.2 - Sample size again

Time spent: 10 hours

5. Other Aspects Related to Methods and Measures Didn't check ▾

Time spent:

II. DATA, CODE, AND STATISTICAL ANALYSES

1. Code Functionality Indeterminable ▾

**Does the provided code run without the need to make any adjustments and without errors?
If not, what steps were needed to get it to run (if it was eventually possible)?**


No code or analysis scripts were shared.

- Note by Recommender: My understanding of the correspondence between the Reviewer and Author is that the data processing and analyses were conducted by a statistician and that the author does not have access to code or scripts to reproduce the results (i.e., not that the Author refused to share).

Time spent:

2. Computational Reproducibility of Reported Statistics Errors found ▾

Is there a clear traceability of reported stats to code? Does the code output match what's reported in the paper? Are all reported statistics findable within the analysis code?

I recreated the data processing and analysis myself. Some results reproduced well (see section  6.1 - Descriptive statistics). But I found many minor discrepancies in the reproduced values for other results:

- 6.2 - Per-question t-tests - Minor discrepancy in a p-value and an effect size

- 7.3 - Averaged t-test - Minor discrepancy in the sign of an effect size
- 10.1.3 - T-test - Minor discrepancy in the sign of an effect size

Time spent: 3 hours

3. Data Processing Errors Errors found ▾

Are there substantive errors during the preparation or cleaning of data (e.g. duplication of rows during a merge) prior to substantive analyses and hypothesis tests?

4.1 - Post-test without pre-test - 2618 subjects had post-test responses without pre-test responses. The issue seems to be a design or coding bug wherein subjects were asked for consent for the post-test even if it was refused in the pre-test. Nevertheless, these subjects were excluded from the original analysis, so there was likely no negative impact.

Time spent: 1 hour

4. Model Misspecification Didn't check ▾

Are there any consequential issues with the assumptions or the form of a statistical model (e.g., overfitting, wrong distribution assumption) used to describe data?

6.5 - Not comparing - The article makes a statistical error by comparing the high p-value of the control conditions with the low p-value of the intervention conditions. However, as explained by Gelman and Stern ([2006](#)), the difference between a significant result and a non-significant result is not necessarily significant. I reran the analyses to test for an interaction effect and found the results to be categorically similar to how the article interprets the original results.

8.2 - Prior susceptibility - The article uses an oversimplified approach that splits the data about the median and ignores the interaction between pre-vs-post and control-vs-intervention. I tested for the interaction using an ordinal ANOVA with results categorically similar to the article's conclusion that people with higher pre-test scores are more likely to see a benefit from the treatment. However, additional analysis raises the possibility that floor and ceiling effects could explain these results.

Time spent: 4.5 hours

5. Erroneous/Impossible/Inconsistent Statistical Reporting

No errors found ▾

Are there inconsistencies between test statistics, degrees of freedom, and p-values? Are there implausible degrees of freedom between compared SEM models? Are there point estimates outside the confidence interval bounds?

[StatCheck Simple Edition](#) did not detect any inconsistencies

Time spent: 5 minutes

6. Other Aspects Related to Data or Code

Didn't check ▾

Time spent:

III. CLAIMS, PRESENTATION, AND INTERPRETATION

1. Interpretation Issues

Indeterminable ▾

Throughout the entire paper, is there an incorrect substantive interpretation of data or statistical tests, causal inference issues, etc.?

7.1 - MANOVA - I was unable to reproduce the MANOVA reported in the article. However, with a lack of sufficient information, it is not clear if I am simply doing the MANOVA a different way. Nevertheless, the MANOVA's results are never actually interpreted in the article, so it is unclear what purpose it serves.

Time spent: 0.5 hours

2. Overclaiming Generalisability

Didn't check ▾

Does the paper overclaim the generalisability of the findings with regards to stimuli, situations, populations, etc.? Is there hyping or overselling of the importance or relevance of findings?

Time spent:

3. Citation Accuracy Errors found ▾

Are there misrepresentations of substantive claims by cited sources? Inaccurate direct quotes? Incorrectly cited or interpreted estimates? Citations of retracted papers?

The article cites ([Lakens 2013](#)) for reporting effect sizes. But it appears to misrepresent the recommended reporting approach. The article reports Cohen's d using the standard deviation of the paired differences (Cohen's d_z), while Hedges' g uses an average of the two standard deviations (Hedges' g_{av}). However, the Lakens article does not recommend reporting Cohen's d_z for this type of study data.

The article only mentions the different pooling methods once. It then reports the effect sizes further down without a subscript that would differentiate them. I was very confused for a couple hours about why I was reproducing everything except Hedges' g .

This issue is primarily a clarity concern. See:

🟡 6.3 - Hedges' g

✅ 7.4 - Hedges' g (clarity issue)

Time spent: 2.5 hours

4. Other Aspects Related to Interpretation Errors found ▾

Time spent:

- Note by Recommender: Reviewer marked this as didn't check but I've changed it to errors found, as the points about 'not comparing' might fit better here within the template.


IV. ADDITIONAL SECTIONS FOR THIS REVIEW

(Additional headings added by the Reviewer)

1. Replicability reporting

Is sufficient code, materials, and information provided that you believe you could replicate the study without help from the authors?

🔴 3.1 - Source code

 3.2 - Exact stimuli

 3.4 - Technical issues

2. Open data

Is the raw data available in a usable documented format?

 3.3 - Data availability

Time spent: 3 hours