ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

ERROR REPORT

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological Psychiatry, 84*(6), 443-451. https://doi.org/10.1016/j.biopsych.2017.12.017

reviewed by

Sascha Meyen, University of Tübingen

July 2, 2024

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.*

I. METHODS, MEASUREMENT, AND DESIGN

1. Design   No errors found

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

Time spent: 60 minutes/hours

There is one aspect of the design that I consider problematic on a conceptual level. However, there was no practical consequences of this problem, so I marked this aspect as "No errors found".
The problem is in combining a staircase procedure (Experiment 2) with measuring M-Ratio: Participants that are volatile in their performance throughout the experiment (producing high accuracies in some blocks and low accuracies in others), will see stimuli that vary substantially in difficulty. This in turn can allow them to better tell apart difficult from easy trials leading to higher values of M-Ratio. In contrast, participants with a stable performance have a less fluctuating staircase sequence keeping stimulus difficulty relatively constant and therefore have it harder to tell difficult from easy stimuli apart leading to lower M-Ratio values. Thus, using a staircase may invalidate M-Ratio interpretations if the staircases produce different variances of the stimulus difficulty for the participants. See Rahnev & Fleming (2019) for details on this problem.

I checked if participants' M-Ratio values are correlated with the variability of the stimulus difficulty, but I found no relevant correlation here. Additionally, I checked whether the variability of the stimulus difficulty was related to the symptom dimensions of the participants. There was also nothing to report. Taken together, the potential conceptual problem did not materialize here.

Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, 2019(1), niz009.

[I generated code for the reproduction which is available here: https://github.com/saschameyen/Error_Reviews_Rouault_et_al_2018_by_Sascha_Meyen

In the following, I will provide the path to the R scripts for my validation analyses in square brackets: Scripts/validation/validate_staircase_procedure.R]

2. Measurement No errors found

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

Time spent: 180 minutes/hours

I am no expert in psychiatric diagnostic so I cannot with high confidence attest to the validity of their use of the screening questionnaires. I read through some of the original studies that introduced the questionnaires (see III 2. Overclaiming Generalisability) and I am somewhat confident that there was no inappropriate application of these measures. From a statistical point of view, their approach is sound. The application of their screenings seems highly sensible. The other methods also did not show any problems.

3. Preregistration Consistency Not applicable

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

Time spent: 0 minutes/hours

There was no preregistration for the study and therefore no deviations.

4. Sampling No 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?

Time spent: 90 minutes/hours

The sampling strategy of Experiment 1 was not explicated. However, this study used a large sample size therefore alleviating concerns regarding statistical power to some degree.

The sampling strategy of Experiment 2 was based solely on the targeted main finding of a correlation between mean confidence ratings and symptom cluster scores (after controlling for age, IQ, and gender). The correlation found in Experiment 1 was r = -.13, which yielded a required sample size of approximately 470 participants. Due to unforeseeably large proportions of exclusion rates in online recruitment, the authors sampled 637 participants and made exclusions such that 497 participants remained in the final sample.

I reproduced the regression coefficient computation from Experiment 1 and the power analysis. There were no errors.

[Scripts/validation/validate_regression.R,
Scripts/validation/validate_power_analysis.R]

5. Other Aspects Related to Methods and Measures Not applicable

Time spent: 0 minutes/hours

There are no other aspects related to methods and measures to be reported.

II. DATA, CODE, AND STATISTICAL ANALYSES

1. Code Functionality No errors found

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)?

Time spent: 30 minutes/hours

The code runs out of the box after installing the relevant R packages. This was a pleasant surprise.

2. Computational Reproducibility of Reported Statistics No 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?
Time spent: 360 minutes/hours

The reported results can be traced back to the data via the freely available methods (https://github.com/metacoglab/RouaultSeowGillanFleming) up to a certain point: What is missing are the early stages of drift diffusion parameter estimation, the metacognitive efficiency (M-Ratio) fits, and the exclusions. Presumably, these computations were left out in the freely available data because they were done in MATLAB by another coauthor than who uploaded the data to github.

I reproduced the perceptual task results and the factor analysis results based on the earliest available data. In this process I noticed that individual trials were excluded for the participants which, I believe, the authors failed to report. However, for the most part, there were very few exclusions (median 1%) making this only a minor oversight. The authors mentioned in personal communications that digging up this data would require more time investment. But given the overall high quality and reliability of their open materials, I believe this is not necessary to make the judgement that the results can be well reproduced. Note that, for this, I also checked that the included data does indeed adhere to the exclusion criteria.

Because the M-Ratio fit was not provided in the freely available materials, I checked reproducibility via different estimation implementations. This yielded very similar results. Altogether, the results reproduced without problems.

Also, the drift diffusion model parameter fit was not provided, and I could not get to successfully run the MATLAB implementation they referred to. I validated their fits by simulating data according to their model. This revealed only a minor issue: Published response times seem to have an offset of 300 ms which is likely due an inconsistent handling for response times — for DDM fits the response time included the stimulus presentation time (of 300 ms) but this offset was removed from the published response times. Beyond that, visual inspection of simulated data based on the parameter fits seem to indicate valid fits and, although this is only a coarse validation approach, I have little doubt that relevant errors occurred here.

[Scripts/validation/validate_factor_analysis.R,
Scripts/validation/validate_m_ratio.R,
Scripts/validation/validate_original_vs_reproduced_data_exp_2.R,
Scripts/validation/validate_perceptual_task_summary_statistics.R,
Scripts/validation/ validate_DDM_fit.R]

3. Data Processing Errors No 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?

Time spent: 120 minutes/hours

No errors of that sort.

[Scripts in reproduce_summary_data_exp_1/ and reproduce_summary_data_exp_2/]

4. Model Misspecification No errors found

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?

Time spent: 120 minutes/hours

Also no errors of that sort.

[Scripts in reproduce_summary_data_exp_1/ and reproduce_summary_data_exp_2/]

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?

Time spent: 60 minutes/hours

The manuscript mentions only estimates (such as regression coefficients) and p-values but no test statistics and degrees of freedoms. However, the provided SEMs in the figures are plausible given the sample size. I ball-parked the SEM for normalized regression coefficients / correlations as roughly 1/sqrt(N). Reanalysis results were true to the original results.

[Scripts/validation/validate_model_comparison.R,
Scripts/validation/validate_regression.R]

6. Other Aspects Related to Data or Code   No errors found

Time spent: 120   minutes/hours

I was wondering why the authors omitted results on M-Ratio for Experiment 1 in their Figure S6A even though they are presented for Experiment 2 in S6B. I reproduced the analysis and obtained very similar results to those of Experiment 2. To me, this alleviates all concerns that a result was intentionally omitted.

[Scripts/validation/validate_missing_metacognitive_efficiency_analysis_in_exp_1.R]

The authors used a logarithmic transformation of the questionnaire data before z-transforming it. Similarly, the authors used a logartihmic transformation of M-Ratio computing log(meta-d'/d'). This log transform seems ad hoc to me. However, I obtained very similar results without this transformation. Therefore, I do not believe there is a problem due to this seemingly ad-hoc rescaling.

Furthermore, I want to mention here that I only detected very minor mistakes in coding or writing. These are barely worth mentioning. For example, the variable names of the Matlab data in Experiment 2 do not match the data because they are copied from Experiment 1. But it is obvious how to repair this by removing one superfluous variable name. Another example is the mentioning of 5 difficulty bins in the Supplement even though there were 6 bins. This is also just a minor typo. Given that this is the level of detail at which mistakes occurred, I am overall very positively surprised with the reproducibility of the results of this study

III. CLAIMS, PRESENTATION, AND INTERPRETATION

1. Interpretation Issues   No errors found

Throughout the entire paper, is there an incorrect substantive interpretation of data or statistical tests, causal inference issues, etc.?
Time spent: 60 minutes/hours

Because the interpretations of their results are mostly straightforward, I feel highly confident that this part (from reported results to interpretation of these results) is solid. In fact, I consider the simplicity of their results a praiseworthy feature of the study.

2. Overclaiming Generalisability No errors found

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: 120 minutes/hours

The weakest point of the study is the reliance on a Mechanical Turk population. The authors make it clear in the supplement that the population they are sampling from is not quite reliable: 13% of participants failed to use the confidence scale properly and had to be excluded (even more based on other sanity checks). It is possible that this population — even after exclusion — deviates in their responses to the clinical questionnaires from the general population. In this case, generalizability would be problematic. The authors did not address this concern in the study, therefore their interpretation of "[o]ur findings indicate a specific and pervasive link between metacognition and mental health" (Abstract) is bordering on overstating the reliability of their results and its generalizability.

To investigate this, I took the samples of the original studies of these questionnaires as a reference as comparison. Overall, the Mechanical Turk sample of Experiment 2 yielded surprisingly similar responses in comparison to these reference samples. Here are three examples:

1. The original authors of the Obsessive-Compulsive Inventory-Revised (Foa et al., 2002) reported that obsessive-compulsive patients have a median score of 25. Patients with other disorders had median scores of 7 and 11. In comparison, the sample of Experiment 2 also had a median score of 11 indicating that the Mechanical Turk population does not respond higher on this questionnaire than expected.

2. The authors of the Short Scales for Measuring Schizotypy (Mason et al., 2005) reported for non-schizotypic participants a mean total score of 13.0 (out of 43 points). The sample of Experiment 2 had a mean score of 12.4, which was again very comparable.

3. For the Zung Self-Rating Depression Scale (Zung, 1965), the clinically relevant group had raw scores above 48 while the sample of Experiment 2 had a mean score of 37, which is in the range of expected values.

Based on these comparisons, there is no evidence that the Mechanical Turk sample of the study deviates from the general population in a meaningful way. Therefore, despite initial doubts, I consider this aspect of the study not problematic.

Foa, E.B., Huppert, J.D., Leiberg, S., Hajcak, G., Langner, R., et al. (2002). The Obsessive-Compulsive Inventory: Development and validation of a short version. Psychological Assessment, 14, 485-496.

Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research*, *78*(2), 293-296.

Zung, W. W. (1965). A self-rating depression scale. *Archives of General Psychiatry*, *12*(1), 63-70.

[Scripts/validation/validate_mturk_sample.R]

3. Citation Accuracy   No errors found

Are there misrepresentations of substantive claims by cited sources? Inaccurate direct quotes? Incorrectly cited or interpreted estimates? Citations of retracted papers?
Time spent: 60   minutes/hours

Some of the cited studies I am aware of, and I found no misinterpretation or incorrect citations within them. However, I want to make clear that I am no expert regarding the clinical aspects of the study and therefore cannot attest to the validity of references to these studies with perfect certainty. But overall, the authors seem to have done a thorough job in their citations, especially in the Supplement and in the online available list of questionnaires. Thus, I found no noteworthy errors here.

4. Other Aspects Related to Interpretation   Not applicable

Time spent: 0   minutes/hours

No other issues are worthy of note.