Introducing the Journal of Robustness Reports

František Bartoš[1⋆], Alexandra Sarafoglou[1], Balazs Aczel[2], Suzanne Hoogeveen[3], Christopher D. Chambers[4] and Eric-Jan Wagenmakers[1]

1 Psychological Methods Unit, University of Amsterdam, Amsterdam, the Netherlands
2 Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary
3 Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands
4 School of Psychology, Cardiff University, Cardiff, United Kingdom

⋆ f.bartos96@gmail.com

Abstract

The vast majority of empirical research articles report a single primary analysis outcome that is the result of a single analysis plan, executed by a single analysis team (usually the team that also designed the experiment and collected the data). However, recent many-analyst projects have demonstrated that different analysis teams generally adopt a unique approach and that there exists considerable variability in the associated conclusions. There appears to be no single optimal statistical analysis plan, and different plausible plans need not lead to the same conclusion. A high variability in outcomes signals that the conclusions are relatively fragile and dependent on the specifics of the analysis plan. Crucially, without multiple teams analyzing the data, it is difficult to gauge the extent to which the conclusions are robust. We have recently proposed that empirical articles of particular scientific interest or societal importance are accompanied by two or three short reports that summarize the results of alternative analyses conducted by independent experts [1]. In order to showcase the practical feasibility and epistemic benefits of this approach we have founded the *Journal of Robustness Reports*, which is dedicated to publishing short reanalyses of empirical findings. This editorial describes the scope and the workflow of the *Journal of Robustness Reports* including the type and format of the published articles. We hope that the *Journal of Robustness Reports* will help make reanalyses of published findings the norm across the empirical sciences.

Received 2025-03-25
Accepted 2025-04-09
Published 2025-04-16
doi:10.21468/JRobustRep.0-Editorial

Check for updates

> Data analysis is like an iceberg, it floats with one-tenth of its mass above water
>
> Adjusted from a quotation incorrectly attributed to Sigmund Freud

In traditional research projects, the main conclusions are usually based on a severely restricted set of statistical analyses. The immediate danger is that of 'model myopia', where researchers fail to test whether their conclusions hold up under plausible alternative analysis approaches. In other words, in traditional research projects it usually remains unclear

whether and to what extent the reported conclusions are either *fragile* or *robust*. Implicitly, researchers may believe that there exists only a single appropriate analysis, or else they may suspect that alternative analyses will yield conclusions that are highly similar. Unfortunately, recent 'many-analysts' projects have conclusively shown that these notions are as convenient as they are false.

Specifically, all many-analysts projects to date have found that different analysis teams tend to adopt unique approaches that may result in substantially different conclusions. For instance, Breznau and colleagues conclude

> "In this typical case of social science research, research teams reported both widely diverging numerical findings and substantive conclusions despite identical start conditions" [2, p.1]

and Silberzahn and colleagues concur:

> "Here, we have demonstrated that as a result of researchers' choices and assumptions during analysis, variation in estimated effect sizes can emerge even when analyses use the same data" [3, p. 351].

The between-analyst variability in conclusions is substantial and has been shown to arise across a wide range of empirical disciplines (e.g., neuroscience, [4–6]; psychology, [3, 7–10]; social science, [2, 11]; medical sciences/epidemiology, [12–14]; biology, [15]; and economics/finance, [16, 17]). Moreover, the variability does not appear to be a result of suboptimal analytic choices [e.g., 2]. This means that for a traditional single-analyst research project, it can be difficult to predict whether or not the conclusions are robust under alternative plausible analysis procedures. The problem of model myopia is exacerbated by the fact that analyses are almost always conducted by the same team that collected the data, opening the door for a biased statistical treatment of the results: statistical cherry-picking that presents the relation between the hypothesis and the data in its most favorable light [e.g., 18].

The resulting state of affairs is visualized in Figure 1 [cf. 19]. The traditional single-analyst research project reveals only the tip of the epistemic iceberg, as the conclusions from other plausible analysis procedures are simply unavailable and remain hidden below the waterline. Moreover, the part of the iceberg that floats above the waterline is not representative of the whole, because it was biased by the fact that the analysis team was deeply involved in the design of the experiment and the collection of the data. In the past, statisticians have focused on the details of the methodological procedure (cf. the discussion between the captain and the sailor in Figure 1) while ignoring the more fundamental problems of model myopia and bias [20]; the captain does not realize that his view on the iceberg is warped and incomplete, and this means that his recommendations, however well-intentioned and statistically sophisticated they may be, amount to nothing more than rearranging the deck chairs on the Titanic.

The cure for model myopia may appear straightforward: conduct a many-analysts study, assess the heterogeneity across analysis teams, and judge the degree to which the qualitative conclusions are fragile or robust [see also 21]. However, many-analysts studies take considerable time and effort to coordinate, and this prohibits their routine application. It has therefore been suggested that results of particularly important studies should be accompanied by two or three short reports that detail the results of alternative analyses [1, 4, 22]. We have referred to these accompanying reports as Synchronous Robustness Reports or SRRs.

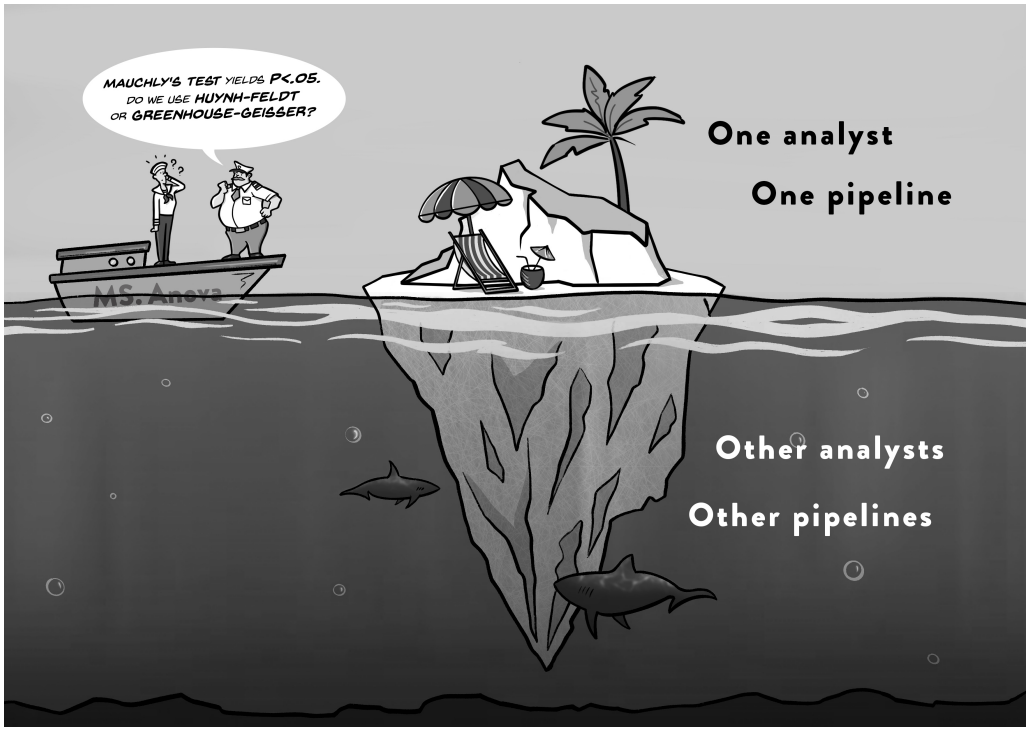

Figure 1: In the traditional data analysis framework, a single data analyst reports the outcome of a single statistical analysis procedure. This framework carries both the risk of model myopia (i.e., only the tip of the epistemic iceberg of uncertainty is visible) and the risk of bias (i.e., the reported analysis may have been unwittingly cherry-picked to provide a flattering impression of the relationship between hypothesis and data; cf. [19]). Figure available under a CC-BY license from BayesianSpectacles.org. Design by Viktor Beekman; concept by Eric-Jan Wagenmakers.

There are various reasons why journals may be reluctant to adopt the SRR format. The format is new, its benefits and feasibility have not yet been demonstrated in practice, and as such the SRR format presents an uncertain proposition. Here we establish the *Journal of Robustness Reports*, a peer-reviewed diamond open-access outlet that aims to showcase the practical feasibility and added value of publishing asynchronous Robustness Reports. The *Journal of Robustness Reports* will spearhead the robustness report initiative and allow readers a more complete and unbiased assessment of empirical results of particular interest. Below we outline the details about journal workflow and the various article types. Most of this information is also available in shortened form on the journal website, https://scipost.org/JRobustRep/about.

JRR article format

The *Journal of Robustness Reports* (henceforth JRR) is a broad, multidisciplinary journal that aims to complement published high-profile empirical findings with concise reanalyses of the original findings using an alternative analysis methodology [i.e., Robustness Reports; cf. 1]. The journal may also offer the original authors the opportunity to respond to the reanalyses, and include an editorial summary that provides an overall synthesis or conclusion.

JRR limits Robustness Reports to only 500 words (excluding the *Acknowledgments and Disclosures* section, the *References* section, figure/table captions, and the title page) and one display element (table/figure), with additional material presented as online supplements.[1] Similar to the SRR guidelines [1], all Robustness Reports in JRR consist of the following seven sections:

1. An *Abstract* section that does not exceed two lines in print. This is equivalent to 180 characters (including spaces).
2. A *Goal* section, which outlines the question that the reanalysis is trying to address.
3. A *Methods* section, which provides background information about the reanalysis.
4. A *Results* section, which describes the main outcomes of the reanalysis.
5. A *Conclusion* section, which compares the results from the reanalysis to those from the original analysis and assesses the degree to which the reanalysis corroborates or undercuts the conclusions from the original analysis.
6. An *Acknowledgments and Disclosures* section, which may consist of the following subsections:
    - A *Reproducibility* subsection. The purpose of this subsection is to declare whether or not the reanalysis team was able to reproduce the original analysis, along with an explanation in those cases where this proved to be impossible.
    - An *Code and Data Availability* subsection. At a minimum, this subsection must include a link to a public FAIR [23] repository that contains the code and data used for the reanalysis. All reanalyses must be fully reproducible. The Robustness Report authors might choose to supply the original findings with additional data, which also need to be made publicly available in the same FAIR repository.[2] The repository may also contain a more extensive version of the Robustness Report.
    - An *Author Contribution* subsection. For multiple-author articles, this subsection may be used to document the contributorship of the different authors [24].
    - A *Funding* subsection.
    - A *Conflict of Interest* subsection.
7. A *References* section.

A constructive response of the original authors [see 25] and the editorial summary are also limited to 500 words and one display element, but need not feature the sections outlined above.

In order to obtain two concrete examples of the JRR article format we first wrote a target manuscript and preprinted it as "People prefer the taste of Belgian mineral water Chaudfontaine over Amsterdam tap water" [26]; next we coordinated two example JRR Robustness Reports (i.e., "Preference for Chaudfontaine or bias towards the preferred option?" [27], "A Bayesian multiverse analysis of the water tasting experiment" [28]) and these now follow this editorial. Although these examples were originally intended merely to illustrate the feasibility of the format, it turned out that they also demonstrated the epistemic advantages of Robustness Reports, as both reanalyses were informative and cast the data in a different light – despite the simplicity of the experimental design.

[1] Extraordinary circumstances may warrant an exception to this rule.
[2] Extraordinary circumstances may warrant an exception to this rule.

JRR article types

As a rule, JRR does not publish any single Robustness Report. A proper assessment of robustness generally requires multiple reanalyses, and therefore JRR requires at least two independent Robustness Reports before either one can be published.

The number of Robustness Reports associated with any particular original article will be relatively low, and therefore it is important that the contributions are coordinated in order to safeguard their methodological diversity and quality. This duty falls to the members of the JRR Editorial College.

Even though prospective authors may suggest an empirical finding for a robustness analysis, and even though they may propose particular analysis teams (including themselves), it is ultimately the responsibility of the JRR Editorial College to invite particular teams to contribute a set of Robustness Reports that is both informative and diverse. Therefore, publication in JRR is by invitation only.

To summarize, submissions to JRR fall into one of four categories:

1. *Robustness Report Inquiry*. An author or team of authors may suggest to contribute a Robustness Report. It is also possible for multiple authors (or multiple teams of authors) to propose a series of Robustness Reports that reanalyze the same data set. If the JRR Editorial College deems the original article as well as the proposed report(s) sufficiently interesting, an official JRR invitation will be issued.

2. *Robustness Report*. The reports are by invitation only in order to ensure statistical diversity and quality.

3. *Invited Reply*. The original authors may be offered the opportunity to respond to the Robustness Reports.

4. *Editorial Perspective*. The editor may decide to comment on the set of Robustness Reports and the Invited Reply in order to provide closure or highlight points of particular relevance.

JRR review process and acceptance criteria

Submissions to JRR have to meet the highest methodological standards and are subject to stringent open peer-witnessed refereeing. To be accepted for publication in JRR, a submission must meet the following expectations and general acceptance criteria (in addition to our standard author obligations):

1. The target article must be of general interest. At the moment when the JRR Editorial College has issued a Robustness Report invitation, this objective is effectively already met.

2. The analyses must be methodologically sound.

3. The contributions must be written in a clear and intelligible way, free of unnecessary jargon, ambiguities, and misrepresentations.

4. The contributions must maintain a respectful and professional tone. Contributions that are characterized by too much 'unnecessary roughness' will *not* be published in JRR.[3]

[3]The same policy applies to reviews, which will be rescinded whenever the comments are insufficiently respectful and professional.

5. The Robustness Report must complement the original analyses of the target article.

6. The analyses must be fully reproducible. The code and data must be shared in a FAIR repository, unless an explicit and compelling reason is provided as to why this either impossible or undesirable.

The outcome of the analysis (i.e., whether a result is statistically significant or not; whether the reanalysis is consistent with the original result or not) is explicitly *not* a consideration for acceptance [cf. 29, 30].

Diamond open access

Publications in JRR are diamond open access. JRR is run by and for the academic community, entirely not-for-profit and without any competing interests. JRR authors retain their copyright: all articles are published in the author's name under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License, allowing freedom to use, reproduce and distribute the articles and related content (unless otherwise noted), for commercial and noncommercial purposes, subject to the citation of the original source. There are no subscription fees, nor are there Article Processing Charges (APCs). By publishing in JRR, authors are helping academic publishing transition to a healthier business model.

Concluding comments

We believe that Robustness Reports hold tremendous promise as a systematic method to encourage a scientific assessment of robustness, incentivize adoption of state-of-the-art statistical methodology, and motivate researchers to showcase the robustness of their conclusions already in the original article [1]. This is particularly relevant for results of profound scientific or societal importance.

The next years will have to show whether and to what extent JRR will flourish. In the best case scenario, Robustness Reports become mainstream and are seamlessly integrated within the current workflow of most existing journals. In this scenario, ultimately there is no more need for JRR: we have founded a journal that aims to make itself redundant. Until then we hope that JRR Robustness Reports can help 'invert the iceberg' of scientific inference and reveal the uncertainty that would otherwise have remained hidden below the waterline forever.

Acknowledgments and disclosure

We are grateful to Nina Ehmann for her assistance.

Funding This work was supported in part by an NWO Open Science grant (OSF23.1.030) and a 2024 Ammodo Science Award.

Conflicts of interest The authors have no financial or proprietary interests in any material discussed in this article.

References

[1] F. Bartoš, A. Sarafoglou, B. Aczel, S. Hoogeveen, C. D. Chambers and E.-J. Wagenmakers, *Introducing synchronous robustness reports*, Nat. Hum. Behav. (2025), doi:10.1038/s41562-025-02129-1.

[2] N. Breznau et al., *Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty*, Proc. Natl. Acad. Sci. **119**, e2203150119 (2022), doi:10.1073/pnas.2203150119.

[3] R. Silberzahn et al., *Many analysts, one data set: Making transparent how variations in analytic choices affect results*, Adv. Methods Pract. Psychol. Sci. **1**, 337 (2018), doi:10.1177/2515245917747646.

[4] R. Botvinik-Nezer et al., *Variability in the analysis of a single neuroimaging dataset by many teams*, Nature **582**, 84 (2020), doi:10.1038/s41586-020-2314-9.

[5] P. Fillard et al., *Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom*, NeuroImage **56**, 220 (2011), doi:10.1016/j.neuroimage.2011.01.032.

[6] K. H. Maier-Hein et al., *The challenge of mapping the human connectome based on diffusion tractography*, Nat. Commun. **8**, 1349 (2017), doi:10.1038/s41467-017-01285-x.

[7] S. Hoogeveen et al., *A many-analysts approach to the relation between religiosity and well-being*, Relig. Brain Behav. **13**, 237 (2022), doi:10.1080/2153599X.2022.2070255.

[8] S. Hoogeveen et al., *Prevalence, patterns and predictors of paranormal beliefs in the Netherlands: A several-analysts approach*, R. Soc. Open Sci. **11** (2024), doi:10.1098/rsos.240049.

[9] G. Dutilh et al., *The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models*, Psychon. Bull. Rev. **26**, 1051 (2018), doi:10.3758/s13423-017-1417-2.

[10] M. Schweinsberg et al., *Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis*, Organ. Behav. Hum. Decis. Process. **165**, 228 (2021), doi:10.1016/j.obhdp.2021.02.003.

[11] M. J. Salganik et al., *Measuring the predictability of life outcomes with a scientific mass collaboration*, Proc. Natl. Acad. Sci. **117**, 8398 (2020), doi:10.1073/pnas.1915006117.

[12] N. N. N. van Dongen et al., *Multiple perspectives on inference for two simple statistical scenarios*, Am. Stat. **73**, 328 (2019), doi:10.1080/00031305.2019.1565553.

[13] Scientific pandemic influenza group on modelling, *SPI-MO: Consensus statement on COVID-19*, UK (2020).

[14] J. A. Bastiaansen et al., *Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology*, J. Psychosom. Res. **137**, 110211 (2020), doi:10.1016/j.jpsychores.2020.110211.

[15] E. Gould et al., *Same data, different analysts: Variation in effect sizes due to analytical decisions in ecology and evolutionary biology*, BMC Biol. **23**, 35 (2025), doi:10.1186/s12915-024-02101-x.

[16] A. J. Menkveld et al., *Nonstandard errors*, J. Finance **79**, 2339 (2024), doi:10.1111/jofi.13337.

[17] N. Huntington-Klein et al., *The influence of hidden researcher decisions in applied microeconomics*, Econ. Inq. **59**, 944 (2021), doi:10.1111/ecin.12992.

[18] T. X. Barber, *Pitfalls in human research: Ten pivotal points*, Pergamon Press, New York, USA, ISBN 9780080209340 (1976).

[19] E.-J. Wagenmakers, A. Sarafoglou and B. Aczel, *Facing the unknown unknowns of data analysis*, Curr. Dir. Psychol. Sci. **32**, 362 (2023), doi:10.1177/09637214231168565.

[20] E.-J. Wagenmakers et al., *Seven steps toward more transparency in statistical practice*, Nat. Hum. Behav. **5**, 1473 (2021), doi:10.1038/s41562-021-01211-8.

[21] A. Sarafoglou et al., *Subjective evidence evaluation survey for many-analysts studies*, R. Soc. Open Sci. **11** (2024), doi:10.1098/rsos.240125.

[22] E.-J. Wagenmakers, A. Sarafoglou and B. Aczel, *One statistical analysis must not rule them all*, Nature **605**, 423 (2022), doi:10.1038/d41586-022-01332-8.

[23] M. D. Wilkinson et al., *The FAIR guiding principles for scientific data management and stewardship*, Sci. Data **3**, 160018 (2016), doi:10.1038/sdata.2016.18.

[24] A. O. Holcombe, M. Kovacs, F. Aust and B. Aczel, *Documenting contributions to scholarly articles using CRediT and tenzing*, PLoS ONE **15**, e0244611 (2020), doi:10.1371/journal.pone.0244611.

[25] C. D. Chambers, R. D. McIntosh and S. Della Sala, *Is 'right-of-reply' right for science?*, Cortex **142**, A1 (2021), doi:10.1016/j.cortex.2021.05.014.

[26] E.-J. Wagenmakers and F. Bartoš, *People prefer the taste of Belgian mineral water Chaudfontaine over Amsterdam tap water*, (PsyArxiv preprint) doi:10.31234/osf.io/2u84v.

[27] A. Sarafoglou, *Preference for Chaudfontaine or bias towards the preferred option?*, J. Robust. Rep. 0-Ex2 (2025), doi:10.21468/JRobustRep.0-Ex2.

[28] S. Hoogeveen, *A Bayesian multiverse analysis of the water tasting experiment*, J. Robust. Rep. 0-Ex1 (2025), doi:10.21468/JRobustRep.0-Ex1.

[29] C. D. Chambers, *Registered reports: A new publishing initiative at Cortex*, Cortex **49**, 609 (2013), doi:10.1016/j.cortex.2012.12.016.

[30] C. Chambers, *The seven deadly sins of psychology*, Princeton University Press, Princeton, USA, ISBN 9781400884940 (2017), doi:10.2307/j.ctvc779w5.