Meta-Psychology, 2023, vol 7, MP.2023.4020
https://doi.org/10.15626/MP.2023.4020
Article type: Guest Editorial
Published under the CC-BY4.0 license

Open data: Not Applicable
Open materials: Not Applicable
Open and reproducible analysis: Not Applicable
Open reviews and editorial process: Not Applicable
Preregistration: No

Edited by: Rickard Carlsson
Reviewed by: Not Reviewed
Analysis reproduced by: Not Applicable

A Plea to Psychology Professional Societies that Publish Journals: Assess Computational Reproducibility

D. Stephen Lindsay[1]
[1]Department of Psychology, University of Victoria

Keywords: editorial, computational reproducibility, meta-psychology

My Bona Fides

I believe that psychology professional societies (e.g., the American Psychological Association, the Association for Psychological Science, and the Psychonomic Society) add value to scientific publishing. I have dedicated many thousands of hours of effort to add to that value. I served as Associate Editor of Journal of *Memory & Language*, Editor in Chief of *Journal of Experimental Psychology: General* (2002-2007), and Editor in Chief of *Psychological Science* (2015-2019). I also served on the Governing Board of the Psychonomic Society and Chaired their Publication Committee when that society transitioned from self-publishing to publishing with Springer (see Lindsay et al., 2023, for an account of that tumultuous time).

I put stock in well-managed peer review. I appreciate judicious stylistic editing and high-quality formatting. I like word-count limits on intros and discussions. I especially prize the curatorial dimension of journal publishing: Scientifically wise filtering through peer review can be a powerful force for good. If the coin of the realm becomes the preprint (aka draft), as promoted by some (e.g., Yarkoni, 2012), I predict many problems. Peer reviewed journals are imperfect, but I believe that they can and often do improve the quality of psychological science (Lindsay, 2020).

Yet journals fall far short of their potential in many ways. Here I argue for one particular potential improvement: Checking the computational reproducibility of statistical analyses before publication (González-Beltrán et al., 2015). It is shocking that in 2023 scientific journals publish empirical articles without checking to see if the analyses are computationally reproducible. That is, if the statistical analyses that the authors report having done are run on the authors' data, are the results the same as those the authors report? Publishing scientific articles without testing computational reproducibility is like selling health remedies without know what's in them.

Putting Your Money Where Your Mouth Is

The American Psychological Association is a major publisher, and many other psychological professional societies publish with giants of the industry such as Elsevier, Sage, Springer, and Wiley-Blackwell. Compared to self-publishing, partnerships with commercial publishers benefit professional societies' journals in multiple ways: wider distribution, larger volume, higher impact, reduced publication lag, higher-quality production, and better web support. Revenue from large commercial publishing partners supports psychological societies in many ways (e.g., funding summer institutes, workshops, travel bursaries, subsidized article processing charges for open-access articles, merit awards, increased internationalization, public outreach for science advocacy, lobbying government decision makers, larger and better-catered meetings, enhanced web presence, etc.). Those benefits may in turn increase membership in the societies. So there is much to be said for partnerships between professional societies and commercial publishers, and many professional societies would face drastic cut-backs and loss of impact without them.

Still, the aims of commercial publishers do not always align with scientific ideals (Brembs et al., 2023; Buranyi, 2017). Publishers create barriers to accessing research, charge shockingly high institutional subscriptions, and may be motivated to hype flashy findings on hot topics and to avoid owning up to errors. The big commercial publishing companies focus on generating income for their shareholders. And they succeed, with very high profit margins. A lot of their income comes from tax payers (who fund schools that pay subscription fees and pay researchers who sometimes pay article processing fees to publish open-access reports of research funded by taxes). Partly for those reasons, early

I thank Tom Hardwicke, Alex Holcombe, Helene Intraub, David Mellor, Daniel Simons, and Simine Vazire for helpful comments on earlier drafts. Flaws are my own responsibility.

career researchers increasingly extol the virtues of free "preprints" (e.g., on PsyArXiv.com) as superior to commercial journals. Academic social media seems awash in scorn for pay-walled journals.

We Have Another Problem

Scientists are ethically obliged to provide their data to other scientists for verification. Data files provided for verification should be complete, accurate, and interpretable. The reported statistical analyses should be reproducible given specification of the same model (Cooper & Guest, 2014).

But psychological scientists often fall short in these regards. Kidwell et al. (2016) found that, among psychology articles explicitly claiming data availability, data were rarely available, correct, usable, and complete. Obels et al. (2020) attempted to reproduce analyses from 62 psychology articles published between 2014 and 2018; for only 21 of those were they able to obtain the data and the analyses and successfully re-run the reported analyses to reproduce the main finding. This is particularly worrisome given that these were all Registered Reports.

Closer to home (for me), Crüwell et al. (2023) reported a study of the first-ever issue of the journal Psychological Science in which every empirical paper had received a data badge. Crüwell et al. (2023) reported that "all 14 articles provided at least some data and six provided analysis code, but only one article was rated to be exactly reproducible, and three were rated as essentially reproducible with minor deviations." Patricia Bauer, who succeeded me as Editor of *Psych Science*, published with this article an Editor's Note in which she correctly noted that criteria for a data badge did not specify inclusion of analytic code, just data. Bauer also noted that Crüwell et al. set a rather high bar in that the scientists who attempted to reproduce analyses did so without any contact with the original researchers. Both of those are valid points, but in my view their import is that a separate badge for analytic code is needed. Happily, just now an open analytic code badge is being introduced.

A few years ago the journal *Cortex* introduced a new article category: Verification Reports (Chambers, 2020). As initially proposed by Sanjay Srivastava (2018), a verification report describes an effort to reproduce the analyses of a published study or to analyze the data of a published study in a new way. The inaugural Verification Report in *Cortex*, by Chalkia et al. (2020), described an arduous journey that began as an effort to replicate a famous experiment published in *Nature* by Schiller et al. (2010). As described in a *Cortex* editorial by McIntosh and Chambers (2020), Schiller signed off on Chalkia et al.'s Registered Report plan for a direct replication. But early in data collection Chalkia et al. noticed that few of their subjects met the exclusion rules. That led them to attempt to reanalyze the original data (which arguably they should have done first – see Larsen, 2020; Nuijten et al., 2016). Quoting McIntosh and Chambers (2020):

> The authors repeat[ed] the critical analyses from Experiment 1, using the exclusion criteria stated in the original *Nature* paper (Schiller et al., 2010), or the criteria stated in the recent addendum (Schiller et al., 2018), or no exclusion criteria, or the idiosyncratic set of exclusions based on qualitative judgements that the original study [reportedly] actually used. Only the last scenario yielded a pattern of results at all consistent with the conclusions of that paper; and even here the critical interaction to test for differences in the reinstatement of fear between groups was not statistically significant.

An episode of the excellent Black Goat (Srivastava et al., 2020) podcast released in August 2020 dedicated 40 minutes to this fiasco and other evidence of errors in the data/analyses of articles published in peer-reviewed psychology journals. The Black Goat hosts bemoaned this state of affairs and speculated as to its origins. It was only in the last minutes of the program that they mentioned the possibility that journals could take responsibility for vetting the computational reproducibility of the claims they publish: "In a decent world, that should be good for the journal too," observed Sanjay Srivastava. Cohost Simine Vazire agreed, saying "Seems like a good way for journals that charge a shit-ton of money . . . to spend that money."

During my term as *Editor of Psychological Science (2015-2019)* serious errors were discovered in several data sets that authors of articles published in *Psych Science* had posted online. These authors knowingly linked their articles to data files that could not withstand scrutiny. Some of these cases ultimately led to retraction (others to Corrigenda). It would have been better for all concerned had the errors been caught before rather than after publication.

Some journals in other fields provide in-house checks on statistical rigour and reproducibility before publication[1]. For example, The Odum Institute for Research in Social Science at the University of North Carolina,

[1] Thanks to David Mellor of the Centre for Open Science for providing most of these examples.

Chapel Hill, employs graduate students to conduct reproducibility checks for the *American Journal of Political Science*. The American Economic Association employs data editors to do similar work for their nine journals. *The British Medical Journal Open Science* has dedicated staff who check adherence to their data-sharing policies. The Center for Open Science currently lists 25 journals that do this. See also *The Journal of Development Economics' Replication Policy.*

A Modest Proposal

Major psychology professional societies that get substantial sums from large commercial presses should, I believe, invest some of that money into ensuring that the analyses they publish are computationally reproducible. Arguably this is a low bar—but it is an achievable one. This is not a matter of policing submissions. Like peer review and copyediting, the aim is to improve scientific quality.

Some might argue that peer reviewers should do this work. Some peer reviewers do look at the data and scripts (as per Richard Morey's Peer Reviewers' Openness Initiative, "Peer Reviewers' Openness Initiative," 2014, September 13). That is laudable. But it is service above and beyond the call of duty of peer reviewers. It is already difficult to recruit reviewers; adding to the demands on them would not help.

I propose that a professional society in psychology that publishes journals conduct a pilot study of the costs and benefits of providing an in-house Stats Adviser (see Appendix for some suggestions as to how to conduct such a feasibility study). If that study indicates that serious problems are rare, then the society might judge that there is no need to check computational reproducibility. If that happens, I will post a video of me eating sweetened desiccated coconut (one of the few foods I despise).

I hereby preregister my prediction that the proposed study would reveal many problems with data files and analysis scripts. Opaque variable names. Missing data. Unreported variables, data exclusions, and transformations. Incorrectly specified models. Scripts that yield results inconsistent with those reported. Scripts that do not run at all. Some of the problems would be minor and easily fixable. Fixing them would be good. Some of the problems would be gutting. Knowing about serious problems before rather than after publication would be good.

Conducting a costs/benefits study would be expensive. To succeed in the role, the Stats Adviser would have to be a highly qualified expert due a good pay rate. And the amount of time required might be substantial. Hardwicke et al. (2018) tried to reproduce statistical analyses of articles published in *Cognition*, and reported that most manuscripts demanded several hours of expert work (some as many as 25 hours, with multiple back-and-forths with the corresponding author). Vilhuber (2019) summarized the considerable challenges of his role as the first Statistical Editor for the American Economics Association. Because the proposed study would be difficult and expensive, undertaking it would likely require the society to reduce expenditures that promote psychological science in other ways.

The proposed feasibility study would also involve non-pecuniary costs. Some authors would be put off by requests to provide data and scripts. The Stats Adviser's input would probably increase editorial lag and it might sometimes add to the action editor's workload. Some authors might dispute the Stats Adviser's assessment. Overall, it would probably be a major pain in the ass. The only way it would be easy is if the Stats Adviser rarely had difficulty reproducing analyses. Finding that out (and being able to advertise it) seems to me to be quite valuable.

Ensuring computational reproducibility would exemplify and amplify the value of the society's journals and help justify their cost. If in-house assessment of computational reproducibility of statistical analyses became standard, it is likely that the non-reproducibility rate would soon plunge[2]. More and more researchers would double check their data files and scripts before submission and use data buddies (Morey & Morey, 2016) and/or workflows to facilitate reproducibility (see Clyburne-Sherin et al., 2018).

A Low Bar

A more ambitious proposal might call for rigorous assessment of the formal appropriateness of the reported analyses (e.g., does the dataset meet the assumptions of the statistical tests?). Some have argued in favour of multiverse analyses. Others have proposed that researchers have multiple independent statisticians make judgments as to how best to analyze the data (Wagenmakers et al., 2021). An argument can be made for also vetting procedural reproducibility (i.e., extent to which the authors provide direct access to information and materials that enable other scientists to replicate a procedure). I can see arguments for all of those approaches, but my proposal is modest.

[2] When Psychological Science first began using Statcheck, about 20% of manuscripts scanned included at least one internally inconsistent inferential statistical test report (e.g., t, df, and p that do not go together). We told authors their work would be scanned by StatCheck and that error rate dwindled to almost nothing

The call here is for a professional society to conduct a short-term pilot study to assess the costs and benefits of providing in-house support aimed at ensuring that when the analyses the authors report having conducted are run on the data the authors report having analyzed, the reported results are reproduced. The goal is merely to ensure that other researchers can take the authors' data and run the authors' analyses and obtain the authors' results.

This is akin to asking a company that sells dietary supplements to conduct independent assays to ensure that the products they sell in fact contain the advertised substances in the claimed amounts. As explained near the outset of this paper, there are reasons to believe that many submissions to psychology journals would not easily pass that test. If it turns out that all or most submissions are easily computationally reproducible, then the society in question could announce that happy outcome and dispense with the idea of providing in-house assessments of computational reproducibility, having shown it to be unnecessary.

A critic might argue that ensuring reproducibility is of little value in and of itself. Indeed, some have argued against efforts toward methodological reform on the ground that what really matters is the development of formal theories (e.g., Devezer et al., 2019; Flake and Fried, 2020; Jamieson and Pexman, 2020; Szollosi et al., 2019; Yarkoni, 2019). These thinkers argue cogently that a narrow focus on cleaving to the rules of statistical inference and prioritizing the replicability of empirical phenomenon cannot, in itself, advance our understanding of psychology. Rocks reliably plummet to the ground when dropped, but establishing that fact tells us little about physics.

If an effect or phenomenon replicates, that does not tell us why. If an effect or phenomenon fails to replicate, that likewise does not tell us why. And what we ultimately want to know is WHY. Computational reproducibility is an even lower standard than replicability, so what I am proposing here cannot, in itself, produce better theories. I do not argue against calls for theory development, difficult as that ambition seems to me to accomplish. But surely journals roil the waters when they publish papers that report findings that cannot be reproduced when the same data are submitted to the same analyses as the original authors reportedly conducted. Reproducibility is not sufficient for progress in scientific psychology, but it seems necessary.

References

Brembs, B., Huneman, P., Schönbrodt, F., Nilsonne, G., Susi, T., Siems, R., Perakakis, P., Trachana, V., Ma, L., & Rodriguez-Cuadrado, S. (2023). Replacing academic journals. https://doi.org/10.5281/zenodo.7643806

Buranyi, S. (2017). Is the staggeringly profitable business of scientific publishing bad for science? https://www.theguardian.com/science/2017/jun/27/profitable-business-scientific-publishing-bad-for-science

Chalkia, A., Schroyens, N., Leng, L., Vanhasbroeck, N., Zenses, A.-K., Van Oudenhove, L., & Beckers, T. (2020). No persistent attenuation of fear memories in humans: A registered replication of the reactivation-extinction effect. *Cortex*, *129*, 496–509. https://doi.org/10.1016/j.cortex.2020.04.017

Chambers, C. D. (2020). Verification reports: A new article type at cortex. *Cortex*, *129*, A1–A3. https://doi.org/10.1016/j.cortex.2020.04.020

Clyburne-Sherin, A., Fei, X., & Green, S. A. (2018). Computational reproducibility via containers in social psychology. *Open Science Framework*, *3(1)*. https://doi.org/10.31234/osf.io/mf82t

Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research*, *27*, 42–49. https://doi.org/10.1016/j.cogsys.2013.05.001

Crüwell, S., Apthorp, D., Baker, B. J., Colling, L., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (2023). What's in a badge? a computational reproducibility investigation of the open data badge policy in one issue of psychological science. *Psychological Science*, *34(4)*, 512–522. https://doi.org/10.1177/09567976221140828

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *14(5)*. https://doi.org/10.1371/journal.pone.0216125

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3(4)*, 456–465. https://doi.org/10.1177/2515245920952393

González-Beltrán, A., Li, P., Zhao, J., Avila-Garcia, M. S., Roos, M., Thompson, M., Horst, E. v. d., Kaliyaperumal, R., Luo, R., Lee, T.-L., Lam, T., Edmunds, S. C., Sansone, S.-A., & Rocca-Serra, P. (2015). From peer-reviewed to peer-reproduced in scholarly publishing: The complementary roles of data models and workflows in bioinformatics. *10(7)*. https://doi.org/10.1371/journal.pone.0127612

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal cognition. *Royal Society Open Science*, *5(8)*, 180448. https://doi.org/10.1098/rsos.180448

Jamieson, R. K., & Pexman, P. M. (2020). Moving beyond 20 questions: We (still) need stronger psychological theory. https://doi.org/https://doi.org/10.1037/cap0000223

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14(5)*, e1002456. https://doi.org/10.1371/journal.pbio.1002456

Larsen, E. G. (2020). *Reproduce before you replicate*. https://erikgahner.dk/2020/reproduce-before-you-replicate/

Lindsay, D. S. (2020). *Journals can and often do enhance psychological science* [Accessed on September 13, 2023]. https://onlineacademiccommunity.uvic.ca/lindsaylab/2020/03/21/journals-can-and-often-do-enhance-psychological-science/

Lindsay, D. S., Ross, B. H., & Hunt, R. R. (2023). *Psychonomic society publications: A participants' account of the transition from self-publishing to partnering with springer* [Manuscript submitted for publication]. https://onlineacademiccommunity.uvic.ca/lindsaylab/wp-content/uploads/sites/4861/2023/09/Lindsay-Ross-Hunt-Psychonomic-Publishing-16-May-2023.pdf

McIntosh, R. D., & Chambers, C. D. (2020). The three r's of scientific integrity: Replicability, reproducibility, and robustness. *Cortex*, *129*, A4–A7. https://doi.org/10.1016/j.cortex.2020.04.019

Morey, R. D., & Morey, C. C. (2016). Habits and open science. *Association for Psychological Science - APS Observer*. https://www.psychologicalscience.org/observer/habits-and-open-science

Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48(4)*, 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science*, *3(2)*, 229–237. https://doi.org/10.1177/2515245920918872

Peer reviewers' openness initiative. (2014, September 13). https://www.opennessinitiative.org/

Srivastava, S. (2018). *In the same spirit as the pottery barn rule. call them verification reports. one-pager online reports linked from the original, where you can report you reran the analyses to verify the numbers in the paper, + possibly ran other analyses to verify strength of conclusions [tweet]*.

Srivastava, S., Tullett, A., & Vazire, S. (2020). *The black goat [audio podcast]*. https://www.theblackgoatpodcast.com/posts/does-not-compute/

Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., Rooij, I. v., Zandt, T. V., & Donkin, C. (2019). Is preregistration worthwhile? *PsyArXiv*. https://doi.org/10.31234/osf.io/x36pz

Vilhuber, L. (2019). Report by the aea data editor. *AEA Papers and Proceedings*, *109*(May), 718–729. https://doi.org/https://doi.org/10.1257/pandp.109.718

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, *5(11)*, Article 11. https://doi.org/10.1038/s41562-021-01211-8

Yarkoni, T. (2012). Designing next-generation platforms for evaluating scientific output: What scientists can learn from the social web. https://doi.org/10.2139/ss.

Yarkoni, T. (2019). The generalizability crisis. *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/jqw35

Appendix

To assess the costs and benefits of verifying computational reproducibility in a professional society's journal articles, I would focus on a journal that is mostly empirical, ethically benign, and analytically straightforward (e.g., for the Psychonomic Society, *Memory & Cog-*

nition). I would recruit a PhD with sophisticated stats/programming chops, wide knowledge of research psychology, excellent communication skills, cultural sensitivity, and a strong record for reliability/punctuality etc. I would pay that person well for a 6-month period. Two months before that person was due to arrive, I would add something like the following to the submission portal for new submissions:

> The X Society is exploring the costs and benefits of providing in-house statistical support with the aim of fostering computational reproducibility (see x for details). If your manuscript is sent for review and the outcome of that initial review is encouraging, then you may be invited to provide (if you had not already done so) de-identified versions of the data files upon which the statistical analyses in your manuscript were based and statistical analysis scripts that would enable an expert to reproduce your analyses. If your manuscript is selected to be part of this feasibility study, then a Statistical Adviser will try to reproduce your analyses. If that person has difficulty doing so, they will work with you to figure out the source of the difficulty. The outcome of this process will be shared with you, the action editor, and, if appropriate, with reviewers of your manuscript, but will otherwise be kept confidential. Please tick one of the options below:
>
> __ Yes, I have or will provide de-identified data and scripts if requested.
>
> __ Sorry, no, I would not do that for some or all of the reported studies for the following reasons: [type explanation here]

Once the Statistical Adviser was in place and manuscripts that meet stats review criteria started coming in, the Adviser would request data and scripts for as many experiments as they could handle[3]. Maybe after three or four months they would have obtained data/scripts from 50 experiments. They would then stop adding new ones and focus on resolving as many of the selected cases as possible over the remaining months of their term.

The Stats Adviser would be instructed not to nit-pick minor matters (especially ones on which folks can reasonably disagree). The Stats Adviser would aim to come across as a supporter and enabler rather than as a cop.

It is worth considering turning this into an experiment, with half of eligible submissions randomly assigned to get support from the Stats Adviser. The outcome measure would be how easily other psychologists could reproduce the analyses in the article.

[3] It might seem more efficient to reserve the Stats Adviser's review for accepted manuscripts. But once a manuscript has been accepted, the social dynamics of the situation shift. I suspect that this would lead to reduced benefits of statistical review.