Dear Psychology Professional Society Journal: Please Check Your Stats

D. Stephen Lindsay[1]

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 hours to efforts 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 the Psychonomic Society transitioned from self-publishing to publishing with Springer.

I put stock in well-managed peer review. I appreciate judicious stylistic editing, high-quality formatting, and rapid global dissemination. I like word-count limits on intros and discussions. I especially prize the curatorial dimension of journal publishing. If the coin of the realm becomes the "preprint" (aka "draft") posted on PsyArXiv (as promoted by some; e.g., Yarkoni, 2012), I predict many problems. Peer reviewed journals are imperfect, of course, but I believe that they can and often do improve the quality of psychological science (Lindsay blog).

In sum, I am a privileged insider. But I am asking psychology professional societies to raise their publishing games. Specifically, I ask that they check the computational reproducibility of statistical analyses before publication (González-Beltrán et al., 2015). It is shocking that in 2020 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 a health remedy without know what's in it.

Putting Your Money Where Your Mouth Is

The American Psychological Association is major publisher, and many other psychological professional societies publish with giants of the industry (e.g., Elsevier, Sage, Springer, Wiley Blackwell). Compared to self-publishing, partnerships with commercial publishers benefit journals in multiple ways (wider distribution, larger volume, higher impact, reduced publication lag, high-quality production, web support, open-access options). Revenue from big publishing houses 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, larger and better-catered meetings, public outreach for science advocacy, enhanced web presence, etc.). Those benefits may in turn increase membership in societies. So there is much to be said for such partnerships, and many 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.[2] 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 aim to generate 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 exorbitant institutional subscription fees and that support researchers who sometimes pay article processing fees to publish open-access reports of research that was itself funded by taxes). Partly for those reasons, early career researchers increasingly extol the virtues of free "preprints" (e.g., on PsyArXiv.com) as superior to commercial journals. Twitter seems awash in scorn for pay-walled journals.

We Have a 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 and the reported analyses should be reproducible. Right?

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.

The journal *Cortex* recently introduced a new article category: Verification Reports (Chambers, 2020). As initially proposed by Sanjay Srivastivas (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

---

[2] Buranyi, Stephen (27 June 2017). "Is the staggeringly profitable business of scientific publishing bad for science?". The Guardian. ISSN 0261-3077; "Time to break academic publishing's stranglehold on research". New Scientist. 21 November 2018. Retrieved 27 November 2018.

should have done first – see [Larsen, 2020](); [Nuijten et al.. 2018)](). 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 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 Srivastivas. 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.[3] For example, the [Odum Institute]() employs graduate students to conduct reproducibility checks for the [*American Journal of Political Science*](). The [American Economic Association]() employs a data editor to do similar work. The [British Medical Journal Open Science]() has dedicated staff who check adherence to their data-sharing policies. See also the policy statement of the [Journal Developmental Economics]().

A Modest Proposal

I don't know how long the current publishing system will last, but 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.

---

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

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). 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 society 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 myself eating this paper.

I predict 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 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.

Granted, conducting a costs/benefits study would be costly. Hardwicke et al. (2018) tried to reproduce 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. The Stats Adviser would be a highly qualified person due a good pay rate. So undertaking the proposed feasibility study would require reducing expenditures that promote psychological science in other ways (e.g., copy editing, conference costs, public outreach).

The proposed feasibility study would also be costly. Some authors would be irked by requests to provide data and scripts. The Stats Adviser's work would sometimes add to the action editor's workload and increase editorial lag. In some cases, 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 worth quite a bit.

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 became standard, it is likely that the non-reproducibility rate would soon plunge.[4] More and more researchers would double check their data files and scripts before

---

[4] 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.

submission and use data buddies (Morey & Morey, 2016) and/or workflows to facilitate reproducibility (see Clyburne-Sherin, Fei, & Green, 2020).

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 (Aczel et al., 2021). I can see arguments for all of those approaches, but my proposal is much more modest.

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 conduced are run on the data the authors report having analyzed, the reported results are reproduced. Simply 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 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 I'm wrong about that – if all or most submissions easily pass the test – then the society in question could simply announce that happy outcome and dispense with providing the proposed service, 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, Nardin, Baumgaertner, & Buzbas, 2019; Flake & Fried, 2019; Pexman & Jamieson, in press; Szollosi et al., 2019; Yarkoni, 2020). 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. 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 it 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. Reproducibility is not sufficient for progress on scientific psychology, but it seems necessary.

Appendix

To assess the feasibility of verifying computational reproducibility in a society articles, I would focus on one of the society's journals that is mostly empirical, ethically benign, and analytically straightforward. I would recruit an advanced grad student or recent PhD with sophistioated stats/programming chops, wide knowledge of research psychology, excellent communication skills, cultural sensitivity, and a strong record for reliability/punctuality etc. I'd pay that person well for a 4-month period. Two months before that person was due to arrive, I'd 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 acceptance or an invitation to revise and resubmit, 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.[5] Maybe after two 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 two months of their term.

Alternatively, the Stats Advisor could work with a subset of submitted manuscript that were sent out for review (rather than only manuscripts that had already been reviewed). That would have less impact on editorial lag. And it would enable the action editor to consider the Stats Advisor's reports at the same time as they considered the reviews. The drawback is that

---

[5] 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.

the Stats Advisor would invest time in manuscripts that were subsequently rejected for other reasons. But that would not be wasted time. The Society would learn from it, and authors could benefit from it.

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.