CAMBRIDGE
UNIVERSITY PRESS

Interdisciplinary lessons and recommendations for the evaluation of replicability in behavioral sciences

Mitch Brown[1] and Donald F. Sacco[2]

[1]Department of Psychological Science, University of Arkansas, Fayetteville, AR, USA and [2]School of Psychology, The University of Southern Mississippi, Hattiesburg, MS, USA
Corresponding author: Mitch Brown; Email: mb103@uark.edu

Abstract

As the scientific community becomes aware of low replicability rates in the extant literature, peer-reviewed journals have begun implementing initiatives with the goal of improving replicability. Such initiatives center around various rules to which authors must adhere to demonstrate their engagement in best practices. Preliminary evidence in the psychological science literature demonstrates a degree of efficacy in these initiatives. With such efficacy in place, it would be advantageous for other fields of behavioral sciences to adopt similar measures. This letter provides a discussion on lessons learned from psychological science while similarly addressing the unique challenges of other sciences to adopt measures that would be most appropriate for their field. We offer broad considerations for peer-reviewed journals in their implementation of specific policies and recommend that governing bodies of science prioritize the funding of research that addresses these measures.

Keywords: replication; research integrity; behavioral sciences; peer review; submission requirements

The replicability of empirical findings in science is disappointingly low. Recent data suggest that 70% of surveyed scientists admit to being unable to replicate another's work (Baker, 2016). Efforts to replicate well-known psychology studies further found only 36% of studies replicated the original findings (Open Science Collaboration, 2015). Nonetheless, the reporting of positive results in scientific literature *increased* by 22% between 1990 and 2007 (Fanelli, 2012). This increase could be partially attributed to "*p*-hacking," an ethically troubling practice of manipulating analyses to shift results into the range considered significant (Simmons et al., 2011). Although some behaviors could be defensible with a priori reasoning (e.g., removing statistical outliers; Sacco et al., 2018, 2019), using these practices to increase the odds of significant results could inflate Type I error rates in published research. Governing bodies of science have gone so far as to call these practices detrimental (NASEM, 2017).

As concerns grew over the prevalence of these practices, various scientific fields have implemented ameliorative systemic reforms. Some academic journals have instituted submission checklists requiring authors to state adherence to best practices (Wicherts et al., 2016). For example, the *Journal of Experimental Social Psychology* requires authors to report every manipulation, measure, and exclusion. With psychology oftentimes leading efforts to develop and implement open science practices (Nosek et al., 2022), evidence of their efficacy exists primarily in this discipline (Brown et al., 2022; Protsko et al., 2023). Recent research has tapped scientists across disciplines with identifying these practices in their respective fields. Scientists in the life sciences report concerns of HARKing, *p*-hacking, selective reporting, and lack of methodological transparency. Political science has concerns with fashion-based selection of research ideas, politicization of research, *p*-hacking, salami-slicing, and selective reporting

(Ravn & Sørensen, 2021; Rubenson, 2021). This letter seeks to document the extent to which journals in behavioral sciences have developed submission requirements to minimize the proliferation of these practices in published research. From there, we provide preliminary evidence for the efficacy of these measures while addressing the practical constraints in more interdisciplinary sciences with tangible recommendations for journals to consider.

Detrimental practices increase the likelihood of a result being deemed "publishable" based on general biases of journals to publish significant findings. Examples of detrimental behaviors include the addition of unjustified covariates into a model on a post hoc basis (Simmons et al., 2011) and selectively reporting findings that support hypotheses (Ioannidis & Trikolinos, 2007). To identify the potential rates of non-replicable findings, research has begun evaluating findings based on their results and calculating estimates of this likelihood. Many of these indices (e.g., p-curves) consider the extent to which p-values cluster around α rather than span the entire critical region of $p < .05$ (Simonsohn et al., 2014), whereas others focus on the probability of results being replicated (e.g., Z-curves; Bartoš & Schimmack, 2022).

As the possibility of assessing replicability increases, an objective set of metrics could start evaluating the efficacy of submission requirements. We have recently begun evaluating these efforts empirically. This endeavor involved calculating p-curves for published findings in major psychology journals (e.g., *Psychological Science* and *Journal of Personality and Social Psychology*) following enactment of submission requirements. We quantified the number of requirements (e.g., reporting all measures and open data), as listed on journals' websites. Journals with more submission requirements had lower estimates of non-replicable findings (Brown et al., 2022). Table 1 provides the list of empirically identified submission rules from this analysis.

As psychology provides an initial model for how to implement best practice policies, other sciences could feel empowered to join this conversation to voice their concerns and needs. The intersection of political and life sciences presents an interesting challenge. Some journals within this purview have begun

Table 1. Examples of submission rules extracted from psychology journals by Brown et al. (2022) to be considered for assessing replicability of published findings

Submission rules for consideration
Disclose multiple tests
Report outliers and exclusions
Report all studies
Report all dependent variables
Report the availability of data/data repository link
Share data (required or recommended)
Follow Journal Article Reporting Standards
Register trials (required or recommended)
Report psychometric properties
Report scoring protocols
Report exact p-values
Report descriptive statistics
Justify choice of mediators
Make all materials available
Make all code available
Provide a file of study materials as presented to participants for reviewers' edification
Report all manipulations
Report power or sensitivity analyses
Report effect size and confidence intervals

Table 2. Examples of recommendations to increase the replicability of published findings in behavioral sciences

Example recommendations
Submission checklists to confirm engagement in best practices
Required reporting of best practices in papers
Open science policies sensitive to the practical constraints of a given field (e.g., proprietary data concerns)
Maintain awareness to which best practices are empirically effective at increasing replicability
Funding from governing bodies of science to develop and assess replicability protocols
Communication with affected parties in the scientific process (e.g., editors, reviewers, and researchers) on the costs and benefits of various reporting rules
Identify optimal submission considerations for interdisciplinary outlets
Develop measures to reduce burdens on editors and reviewers during peer review

implementing submission requirements (e.g., *Evolution and Human Behavior* and *Political Psychology*). Conversely, *Politics and the Life Sciences* uses a version of these policies to encourage transparency and best practices but not as a requirement. This discrepancy could reflect a relatively moving target in interdisciplinary sciences based on constraints in their field. Outlets with less explicit ties to psychology may have different criteria for reporting results that could make the implementation of a standardized battery of requirements for these journals difficult. Journals in these areas could nonetheless begin comparing outlets with and without submission requirements. Even without reported *p*-values, many results remain amenable to analyses (e.g., confidence intervals). For qualitative analyses, researchers and journals could collaboratively develop metrics to assess robustness appropriate for the methodology.

As outlets in political and life sciences implement similar policies, it remains advantageous to consider collaborative discussions among those in the peer review process. Objective analyses of submission requirements could be complemented by system-level feedback from authors, reviewers, and editors. This feedback could inform policy based on what requirements could be helpful while identifying various burdens of these requirements and how to address them. For authors, for example, requirements could be prohibitive without special permission (e.g., proprietary data and participant privacy), which may increase systemic barriers for early-career researchers or those at smaller institutes (e.g., Beer et al., 2023; Begum Ali et al., 2023; McDermott, 2022; Mulligan, 2013; Rubenson, 2021). Nonetheless, such measures could prove popular with reviewers and editors. Based on actual reviewer responses in MEDLINE in 2015, 63.8 million hours were dedicated to peer review; 20% of reviewers performed 60%–94% of reviews (Kovanis et al., 2016). This suggests that peer review can be burdensome. Requirements could allow editors to screen submissions and vet them more easily before sending them for review (i.e., desk rejections). Reviewers could provide more substantive reviews efficiently without needing to parse ambiguous findings that may obfuscate detrimental behaviors.

The increasing need for transparency in sciences requires governing bodies to address the appetite of participants to engage in best practices. In addition to rewarding engagement, research could begin investigating how to increase participation and reduce barriers to participation. Such efforts may require funding from outlets in political and life sciences, but governing bodies would benefit from putting their money where their mouths are given how effective they appear to be increasing empirical rigor in behavioral sciences (Protsko et al., 2023). Table 2 provides a summary of the potential recommendations discussed in this letter.

References

Baker, M. (2016). Reproducibility crisis. *Nature*, **533**, 353–366.

Bartoš, F., & Schimmack, U. (2022). Z-curve 2.0: Estimating replication rates and discovery rates. *Meta-Psychology*, **6**, 1–14.

Beer, J., Eastwick, P., & Goh, J. X. (2023). Hits and misses in the last decade of open science: Researchers from different subfields and career stages offer personal reflections and suggestions. *Social Psychological Bulletin*, **18**, 1–23.

Begum Ali, J., Holman, R., Goodwin, A. L., Heraty, S., & Jones, E. J. (2023). Parent attitudes towards data sharing in developmental science. *Open Research Europe*, **3**, 182.

Brown, M., McGrath, R. E., & Sacco, D. F. (2022). Preliminary evidence for an association between journal submission requirements and reproducibility of published findings: A pilot study. *Journal of Empirical Research on Human Research Ethics*, **17**, 267–274.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, **90**, 891–904.

Ioannidis, J. P., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal*, **176**, 1091–1096.

Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016). The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLOS ONE*, **11**, e0166387.

McDermott, R. (2022). Breaking free: How preregistration hurts scholars and science. *Politics and the Life Sciences*, **41**, 55–59.

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology*, **64**, 132–161.

National Academies of Science, Engineering, and Medicine. (2017). *Detrimental research practices*. National Academies of Science, Engineering, and Medicine.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M.K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, **73**, 719–748.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, **349**, aac4716.

Protzko, J., Krosnick, J., Nelson, L., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., & Schooler, J. W. (2023). High replicability of newly discovered social-behavioural findings is achievable. *Nature Human Behaviour*, 1–9. https://doi.org/10.1038/s41562-023-01749-9

Ravn, T., & Sørensen, M. P. (2021). Exploring the gray area: Similarities and differences in questionable research practices (QRPs) across main areas of research. *Science and Engineering Ethics*, **27**, Article 40. https://doi.org/10.1007/s11948-021-00310-z

Rubenson, D. (2021). Tie my hands loosely: Pre-analysis plans in political science. *Politics and the Life Sciences*, **40**, 142–151.

Sacco, D. F., & Brown, M. (2019). Assessing the efficacy of a training intervention to reduce acceptance of questionable research practices in psychology graduate students. *Journal of Empirical Research on Human Research Ethics*, **14**, 209–218.

Sacco, D. F., Bruton, S. V., & Brown, M. (2018). In defense of the questionable: Defining the basis of research scientists' engagement in questionable research practices. *Journal of Empirical Research on Human Research Ethics*, **13**, 101–110.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, **22**, 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, **143**, 534–547.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, **7**, 1832.