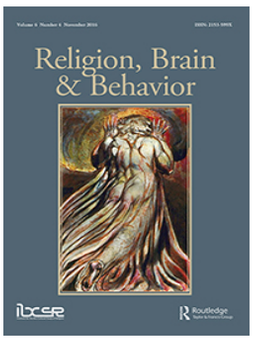

Routledge
Taylor & Francis Group

Religion, Brain & Behavior

ISSN: 2153-599X (Print) 2153-5981 (Online) Journal homepage: http://www.tandfonline.com/loi/rrbb20

Standards for Publishing in Religion, Brain & Behavior

Joseph Bulbulia, Michael L. Spezio, Richard Sosis & Wesley J. Wildman

To cite this article: Joseph Bulbulia, Michael L. Spezio, Richard Sosis & Wesley J. Wildman (2016) Standards for Publishing in Religion, Brain & Behavior, Religion, Brain & Behavior, 6:4, 275-277, DOI: 10.1080/2153599X.2016.1227123

To link to this article: http://dx.doi.org/10.1080/2153599X.2016.1227123

RELIGION, BRAIN & BEHAVIOR, 2016
VOL. 6, NO. 4, 275–277
http://dx.doi.org/10.1080/2153599X.2016.1227123

Routledge
Taylor & Francis Group

EDITORIAL

Standards for Publishing in Religion, Brain & Behavior

Here, we clarify *Religion, Brain & Behavior*'s response to the "replication crisis" in the human sciences. Our purpose is not to reflect on the replication crisis as such, but rather to use it as an entry point for describing standards to which *RBB* holds its data-science submissions accountable.

The term "replication crisis" has different meanings. When used narrowly, the term denotes a syndrome, the widespread failure of researchers to independently reproduce the results of other researchers. The upshot: merely because a study has passed peer review does not mean its findings are reliable. The worry: failures to replicate undermine confidence in the human sciences.

When used broadly, the term "replication crisis" offers a diagnosis for the failure-to-reproduce syndrome. Diagnoses vary. However, critics agree that, at bottom, the crisis arises from pressures to publish. From the moment human scientists undertake research they face competition. Graduate admissions are limited, postdoctoral fellowships are rare, rarer still are tenure-track jobs; tenure is denied; promotion is denied; media spots are limited; the funding that enables research is fiercely competitive. The quality and quantity of published research is the most important factor deciding academic fates. Top journals, too, face competition. Research that is highly cited and highly publicized increases a journal's impact factor. The most cited journals explicitly look for attention-grabbing results. The drive for shocking findings creates incentives for researchers to find them, regardless of whether results reflect reality. Though few researchers fabricate data, critics urge that pressures to publish have compromised the human sciences by incentivizing remarkable over accurate research.

Critics have suggested various proposals for dealing the replication crisis:

Pre-registering hypotheses. Stating a hypothesis in advance of conducting a study is thought to restrict researcher degrees of freedom when analyzing the data. Pre-registration is meant to prevent the practice of collecting lots of data and then fishing for results.

Meta-analysis. By pooling the findings of many studies, including unpublished studies, researchers obtain larger samples, which are informed by failures to replicate, enabling more accurate inferences.

Increasing sample size. Larger samples augment power in detecting effects, and overcome sampling biases. Not only are small studies prone to error, errors can run in the opposite direction to true effects.

Reporting effect sizes conveys an estimate of the magnitude of differences between treatment/exposure conditions and control conditions. Reporting effect sizes has the advantage of clarifying whether differences are reliably large or small.

Reporting confidence intervals clarifies the range values an estimated effect might take. Confidence intervals are sometimes promoted as an alternative to classical p-values, which are statistics that describe the likelihood of an effect at least as extreme as the observed effect.

Bayesian data analysis has been promoted as an alternative to classical null-hypothesis significance testing. Bayesian estimation recovers natural posterior probabilities for estimated effects, conditional on the data at hand, the statistical model, and any prior information researchers might include about the mechanisms that give rise to the data (Kruschke, Aguinis, & Joo, 2012). Bayesian estimation is computationally intensive, so faster computers and the growth of open-source statistical software has rendered Bayesian modeling widely accessible.

There is no shortage of advice about how to handle replication failures. Each of the suggested methods has its limitations.

Preregistration does not prevent fishing. Researcher degrees-of-freedom abound even when hypotheses are preregistered (Gelman & Loken, 2014). Additionally human scientists often make discoveries by examining and exploring their data. Banishing exploration may compromise discovery.
Meta-analysis is a powerful scientific tool but it is inherently limited by the questions researchers have asked in the past, and by the limitations of the studies that have been published or filed away. A meta-analytic harvest relies on the seeds of original research.
Increasing sample sizes will improve a study's power to detect an effect. However, because there are rarely *no differences* between two exposures, increasing sample sizes also reduces p-values, making it easier, not harder, to obtain "statistically significant" results. This is a worry if people confuse statistical and practical significance.
Effect size statistics can enter into decision-theoretic frameworks but the practical interest of a result may be poorly reflected by an effect-size statistic. Summing over a population, small effects might lead to substantial human benefits or harms. Large effects might not be interesting.
Confidence intervals provide useful information about the range of likely effects but widespread use of 95% interval as an arbitrary make-or-break threshold does not improve on the use of p-values under .05 as measure of a study's truth.
Bayesian estimation with non-informative priors recovers frequentist estimates. Good estimation hygiene cannot redress poorly measured responses, poor research designs, and default-thresholds humbug. There are no magic Bayesian bullets either.

So how shall *RBB* editors and reviewers evaluate submissions? Shall we require a conjunction of these practices? No.

We reject the very idea of default standards for judging the acceptability of a scientific report. Arbitrary rules and thresholds, the use of methods and models, and the reporting of test statistics cannot replace careful and honest reasoning about scientific and practical inference from a study. The crisis in the human sciences is not grounded in failures to replicate, but rather in adoption of an all-or-nothing mindset about scientific progress. Relatedly, we reject the premise that any published research should be regarded as true merely because it has been published. This includes published research using preregistered hypotheses or published research that replicates previous findings. More specifically, we reject an all-or-nothing conception of scientific publishing according to which results that meet a standard are true and others are false. Such an all-or-nothing conception of science poorly reflects the history of science. Science accumulates understanding progressively by revising and often reversing prior beliefs. We emphatically advocate practices of vigorous peer review and urge the wider pursuit of replication because surviving peer review and replication tests increases confidence in the model under scrutiny. However, scientific positions are inherently shaded by degrees of confidence.

By what criteria shall *RBB* referees and editors assess data-science submissions?

- We will evaluate all data-oriented studies based on their designs and their questions, not any specific result.
- We require that authors clearly state the question their study hopes to address. What do the researchers want to better understand? No data-oriented study will be considered publishable unless its research question is made crystal clear.
- We require that authors clearly describe the intellectual motivations for addressing their question. All submissions will be presumed unworthy of attention until those intellectual motivations are clarified.
- We require that researchers clearly describe research protocols in careful detail, including measurement methods and instruments. Such statements may be best presented in on-line supplemental materials but they cannot be omitted. This should include a brief discussion of the limitations of protocols and measurement methods.
- We strongly recommend that, unless prohibited for ethical or legal reasons, authors make their data publicly available for the purposes of replication, along with explicit and detailed modeling protocols (such as R scripts.)

- We ask that authors make effective use of graphical presentations of results and inferences.
- A finding's scientific and practical importance (if any) must be separately and clearly stated.
- We require that researchers who work within a null hypothesis significance testing paradigm (NHST) do not confuse failure to reject the null hypothesis as acceptance of the null hypothesis. Though we will not exclude NHST reasoning, we'd prefer that people gave up the idea of "acceptance" and "rejection" of a hypothesis altogether and return to thinking about how a finding affects uncertainty about the question at hand. Minimally, we will not tolerate claims that obtaining a specific test statistic (e.g. $p<0.05$) renders a finding "significant."
- One of our key interests in evaluating a data-oriented study will be in how it quantifies uncertainty around the question it addresses. How much more or less should a community of interested researchers be confident about the study's results based on the data collected and its analysis?
- We ask that researchers clarify how they attempted to make their paper's result disappear. Authors who clarify how their result was destroyed will not be penalized. Quite the opposite: such authors will receive our quiet praise and the respect of their peers.

This partial list of standards is neither exhaustive nor fixed. As knowledge grows, our standards of accountability will change. Though we expect our rejection of default test-statistic thresholds (p-values, confidence intervals, effect sizes) will come as a relief to aspiring *RBB* authors, the principles we have identified here make it more difficult to publish with us, not less. P-values can be hacked. Test statistics can be fudged. Grandiose pronouncements are much easier than balanced, sober assessments of one's results. By removing exclusive reliance on test-statistics to establish value, researchers will have nowhere to hide. To publish in *RBB,* data-science authors will need to think harder about the scientific and practical significance of their findings, and will need to plainly communicate that significance to a scholarly community.

References

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *Am Sci*, *102*(6), 460.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722–752.

Joseph Bulbulia
Michael L. Spezio
Richard Sosis
Wesley J. Wildman