Revisions for Measuring Transparency in the Social Sciences: Political Science and International Relations

Royal Society Open Science Manuscript ID #RSOS-240313

Bermond Scoggins and Matthew P. Robertson

We greatly appreciate the opportunity to revise the manuscript and thank Dr. Hardwicke and the two reviewers for the time they have spent with the paper and providing helpful comments and guidance. In incorporating reviewer feedback, we believe we have improved the manuscript's quality.

Below, we outline point-by-point responses to each of Dr. Hardwicke and the reviewers' comments and identify where changes have been made in the manuscript.

Dr Tom Hardwicke

(1) *I want to draw attention in particular to the comments regarding transparency of methods — the authors must ensure that the methods are described in comprehensive detail. The level of detail expected should enable an independent researcher to repeat the authors' methods and obtain the same results.*

(2) *At least one reviewer had some trouble identifying and interpreting the validation process/results — please can the authors clarify this.*

(3) *One reviewer suggests providing policy prescriptions — it is up to the authors whether they follow this suggestion, but from my perspective I do not think it is necessary in an empirical paper that is not specifically evaluating the policy being advocated for. If the authors do offer policy suggestions, they should make clear whether their suggestions are directly evidence-based or merely speculations inspired by the evidence.*

We believe that, given the descriptive scope of our data and analysis, we are not well positioned to provide policy prescriptions. We identify that the journals whose editors were signatories to the Data Access and Research Transparency (DA-RT) statement have followed through and

open data rates have increased to their intended levels. While very important, we cannot speak to how to incentivise non-DA-RT journal editors to adopt the reforms the DA-RT statement advises.

(4) *Please ensure that empirical claims in the introduction are supported with citations. For example: "Yet current assessments of the field's progress have been based on relatively small samples".*

We have corrected this and added several citations to published PSIR data availability papers. These papers all use small samples and rely on time-intensive human coding procedures. We have included a footnote outlining how these papers have limited spatial and temporal coverage: They examine, at most, a few hundred papers drawn from a small number of top journals and cover only one or two years.

In line with reviewer 2, we have also clarified in the introduction that open science practices have been advocated for decades. We have cited our discipline's Gary King who wrote a paper in 1995 advocating for replications (i.e. open data where computational reproducibility can be tested and analytic extensions conducted).

(5) *The paragraph starting "In 2020, the percentage of statistical inference papers with open data in political science journals was approximately…" appears to report results from the present study in the introduction section. Results should not be reported in the introduction, only in the results and discussion section.*

The discussion of the results in the introduction have been removed.

(6) *Should headings 2, 3, and 4, be sub-headings of the introduction section?*

We agree and have made these sections sub-headings of the introduction.

(7) *Note that the term "replication" has different meanings in different fields so please provide an explicit definition.*

This is a good point and we have changed the terminology throughout the paper to clearly distinguish between replication – defined as collecting or sampling new data to test the same theory or hypotheses as the original paper – from computational reproducibility – being able to access the data and code from the original paper to determine whether the code reproduces the same output as published. Since we refer extensively to the replication crisis, primarily in psychology, we now clearly distinguish between replication and reproducibility despite political science usually referring to the latter as "replicibility". A paragraph has been added in the introduction defining both open science practices.

(8) *I see that code and data have been shared on Github which is great, however, Github is not considered a trusted third party repository (see e.g., https://social-science-data-editors.github.io/guidance/Requested_information_hosting.html#trusted-repositories).*

To ensure the accessibility and longevity of this important information, I suggest the authors link the Github repo to OSF or Zenodo and generate a DOI link.

Reviewer 1

(9) *The open science movement did not begin "in the 2010s". See the "Replication, Replication" symposium in 1995 in PS, among others.*

(10) *Preregistration is not necessarily better; it depends on the nature of the study. You preregister if you can't run new analyses very easily, but should not preregister if the goal is to discover new questions, new research directions, and serendipity. You need to clarify and classify articles before deciding whether lack of preregistration is good or bad. You say "As we show below, preregistration is not yet the norm in political science and international relations." I would guess that if it had been the (rigidly enforced) norm for the last few decades, most progress in the field would not have been made.*

(11) *Replicability these days varies by journal more than author. Can you report journals? There's little point in checking individual articles for journals that require as a condition of publication replication data, especially including some that replicate the results before final acceptance.*

(12) *Many, and I'd wager most, social science methodologists would recommend that test statistics not be used at all, and in its place point estimates and confidence intervals. Many of the problems the authors note with their rules would vanish if the authors were focused on the quantity of real interest rather than some arbitrary statistical criterion that does not necessarily translate into substantive problems of interest.*

(13) *"Many Labs studies (2014, 2018) have shown that the effect sizes in highly powered replications are much smaller than those in the original studies." So who would you trust? The original authors – who often spent many years working on their papers – or the replicators – who typically spend an hour or two on a replication? Why is the replicator the judge? That's not how science works, nor should it.*

(14) *The first 11 pages of this manuscript could be reduced. The paper really starts now on p.12.*

(15) *Dictionary based text analysis is 3-4 scholarly generations of text analysis behind the cutting edge. This should be replaced with modern automated text analysis methods, maybe even LLMs. But whatever you use (perhaps even dictionary based methods), the authors must have some validation that their method does what they claim.*

(16) *The ML models are also generations behind current best practice. Ensembles, random forests, deep learning, and others routinely beat SVM and NB. Percent accuracy is not relevant, unless they are 100% accurate. We need evidence on bias and variance. The authors talk about confusion matrices (which is the right thing), but I don't see the Appendix 1 they refer to. And we also need to see what kinds of corrections were made on the basis of these.*

We thank the reviewer for pointing out the omission of confusion matrices in the appendix. We have added the SVM confusion matrix, which is the classifier we used to perform our main analyses. The figure also includes the main statistics during the train-test process. These statistics indicate the model performs very well and particularly in identifying statistical papers with 94% sensitivity.

Reviewer 2

(17) *The authors note that previous efforts to assess the availability of open data have been based on "small samples." But even more importantly, these small samples have been unrepresentative, mostly focusing on the "top" or "most visible" journals. The authors offer a careful discussion of the literature on reproducibility and replication. This review is thorough, perhaps unnecessarily so. If the authors are limited by space, they might consider condensing this material.*

We appreciate this feedback and agree that the first half of the paper can be condensed. We have specifically reduced the length of the open data and preregistration sections.

(18) *As with very complex computational projects, it is difficult to fully describe the procedures in the text, and not everything can or should be fully described. However, to me, the discussion of the procedures seemed slightly on the sparse side. For example, what do the authors mean by "statistical inference paper"? The definition is a little vague. Examples: What CrossRef metadata do the authors use? How complete and accurate is the published.print variable? How does CrossRef record the published.print variable for online-only journals? When the PDF is converted to plain text, how are URLs preserved? How accurate is the conversion? How many of the PDFs require OCR? Why don't the authors use HTML? What do the authors mean by "UNIX command line utilities" and OCR software? What procedure did the authors follow to revise dictionaries and evaluate discrimination?*

(19) *In Section 5.2, the authors describe the critical step of linking papers with their associated data. I think I understand how this works if the journal has a dataverse and the authors post there (steps 1 through 4). But how does it work if the authors share their data somewhere else, perhaps their personal Dataverse? It seems like steps 5, 6, and 7 describe this, but this is pretty sparse for such a seemingly critical procedure. I don't have a good sense of how the authors could know if this was working poorly.*

(20) *As I read step 1 of the procedure to identify shared preregistrations, it seems like the authors only search for "pre-reg" or "prereg." Surely there must be other relevant strings to detect wordings like "pre-analysis plan"?*

This is a good point and we further detailed in that section that we also searched for "pre-analysis" and "pre-analysis". The `count_prereg_papers.R` file details the different terms we searcher for. These include full names and links to the main repositories – Open Science Foundation/osf.org, Evidence in Governance and Politics/egap.org, aspredicted/aspredicted.org.

(21) *Why do the authors split IR and political science into separate analyses? This is a little weird since many journals publish papers in both, so I think there's substantial overlap between the two datasets.*

This was a very helpful comment. We agree that many journals publish both types of papers and that it is conceptually difficult to distinguish the two. We have changed Figure 1 and removed the PS and IR labels – we now report the proportion of all statistical inference papers with a linked open data repository in each year.

We initially followed the journal categorisation labels in the Journal Citation Report 2020 (JCR), which are listed in Appendix Table 1. However, these do seem arbitrary. After scrutinising the list more closely, we know that journals coded as solely PS or IR publish articles belonging to the other category.

(22) *Overall, the results section seemed a little sparse. It's hard for me to make a concrete suggestion here, other than I expected a lot more detail. The authors have collected a rich, comprehensive dataset; it seems at least worth plotting the time series for each journal, perhaps after dividing the 170ish journals into several categories to prevent over-plotting. It also seems worth describing in a lot of detail the status quo—what does the availability rate look like across the 170ish journals for the most recent year in the dataset? How many have no replication data available? How many are above 50%? How many above 90%? These are ideas, but the authors will have much better ideas of what their data can show.*

(23) *Lastly, I think the authors are in a good position to make policy prescriptions. What can political science do to improve? What should authors do? What should journals ask, encourage, and require? What infrastructure should we use and/or develop? What can other disciplines learn from political science?*