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 we thank Dr. Hardwicke and the two reviewers for the time they spent with the paper and the very helpful comments and guidance they gave. We think the feedback has significantly improved the work. We also appreciate the time extension for this response.

Below are our responses to each of Dr. Hardwicke and the reviewers' comments, as well as descriptions of the corresponding changes to 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.*

Indeed. Our challenge was to balance the sheer complexity and tediousness of the analysis (at current count we now have 45 `.R` files, 4 `.py` files, various bash scripts, and over 6,000 lines of code), the limits of word count, and reader patience, against transparency and thoroughness. We've now given a more detailed account of our process, and have supplemented this with an even more detailed account of the specific code files, and their inputs and outputs, in a README file for other scholars. Importantly, we've added additional footnotes pointing to the original code files that contain the analysis we describe in each step. These files have also been supplemented with several paragraphs of description of what each file does. Due to copyright we are unable to supply the full text of the files that are necessary to actually reproduce our analysis, but we provide every other piece of information necessary to obtain those files and reproduce it.

(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.*

Given that this is a descriptive paper, we indeed prefer to hold off on policy recommendations for the time being. 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. We continue to think about the types of policies and practices that would encourage and facilitate transparent research practices, and we have some more work planned in this area.

(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 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, citing King (1995).

(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 has been removed.

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

We agree and have made these 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.*

This is a great point and we have remedied this by using Zenodo's built-in GitHub preservation feature. This has added a DOI and a badge to the repo and it is now preserved by Zenodo.

Reviewer 1

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

An unfortunate oversight on our part. We have fixed this.

(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.*

We have reduced the length of the manuscript considerably.

(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.*

We spent a great deal of time addressing this piece of feedback, going some way to building out an LLM classification pipeline, but in the end abandoned it due to time and resource constraints. A more detailed account of this follows below. However, in the end it made most sense to stick with the SVM model we already have in the paper. We've now added a confusion matrix and further explanations. We do not think that the age of the technique is the most important factor (Gauss discovered regression in 1795), but its accuracy and suitability for the task at hand. We have expanded on this in the paper, and we hope the confusion matrix helps.

Our attempts at using an LLM classifier

The repo now contains a folder `./llm_classifier` that contains the code, data, and preliminary model results for our abortive LLM pipeline.

We used Anthropic's `claude-3-haiku-20240307` model, for three reasons:

1. Very strong performance on benchmarks;
2. Cost ($0.25/million tokens input, $1.25/MTok output, or about half of gpt-3.5-turbo);
3. Speed/accessibility/time constraints[1]

Our process was as follows:

- We did some experimental tests on the terminal using the Anthropic API, then moved this to a python script after confirming functionality.

[1]We don't have the hardware or time to fine tune an open source model. We experimented with the Llama 3 instruct model locally via the `ollama` CLI (details: 8B parameters, 4-bit quantization, ID number a6990ed6be41 available at https://ollama.com/library/llama3) but found the latency far too high and performance far below what we needed.

- We selected a random sample of 100 papers from the pool of 1,624 previously hand-coded papers, which included both papers that use statistical inference (stat_bool = 1) and those that do not (stat_bool = 0). Because the majority of the rows are 0s, to ensure a sufficient representation of papers using statistical inference we disproportionately allocated 70 samples to papers with stat_bool = 1 and 30 samples to papers with stat_bool = 0. This stratified sampling approach allowed us to assess the model's performance with an emphasis on its sensitivity (recall) in the first stage.
- We then began experimenting with the prompt, manipulating a variety of parameters in an attempt to optimize both classification accuracy and resource use. Because we have around 1.5 billion tokens of text in our corpus, and we are paying for the model inference personally, we sought to save money by truncating parts of the article and sending it to the model with a prompt telling the model to perform a classification.
- We found that in the 100 results, the model's answer differed from our own 'ground truth' classification in nine cases. Upon examination, however, we found that our own classifications suffered measurement error – in several of them, `claude-3-haiku` was right and our initial classifications were wrong (though for subtle reasons that required close examination of the papers). In others we tweaked the prompt to get an accurate response (for instance, truncating less of the paper, or starting the truncation at 40% of the paper rather than the beginning, or explicitly mentioning a wider variety of statistical methods). We corrected the mistakes in the hand-coded dataset and reran the model on the 1,624. However, we found the model got 375 wrong, an accuracy of only 76.9%, far below our support vector machine featured in the paper.
- While it would be theoretically feasible to either further experiment with `claude-3-haiku`, or use a better (and more expensive) model, or fine tune either a commercial or open source foundation model, we unfortunately do not have funding for either the inference or GPU compute, nor did we have time to further prove out an LLM for the present project.

Generally, we agree that LLMs are going to be the superior method of binary classification and structured data extraction for tasks like this. We look forward to leveraging these methods on this corpus under fewer resource and time constraints in the future. At present, we feel that a classification accuracy of 92.35% is adequate and we hope the reviewers agree.

(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

analysis. 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 92.35% accuracy and 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 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?*

As mentioned above, we have expanded the discussion of the methods in the paper, hopefully addressing these questions and concerns. We hope the extended length will not pose a problem. A clear definition of a "statistical inference paper" has also now been provided. Where the explanation was thought to be tedious we kept in brief in the paper, added a footnote, and pointed interested readers to a code file that demonstrates the idea or discovery.

Some issues we did not address in the paper, for reasons of space and focus:

- *"How does CrossRef record the published.print variable for online-only journals?"*

Frustratingly, we did not think of this at the time. Crossref does not have a category for online-only. However, one may infer that a journal is online-only if it has zero or very few entries for its print publications. When initially coming up with the criteria, we looked closely at the fields and pulled random samples from the data frames. It became clear that *none* of the other fields were accurate measures of publication date. And we didn't notice any journals

that had *no* entries for print publication date. Indeed, it was most often the most complete field.

But your question triggered a deeper investigation of our dataset, and we discovered that we were indeed filtered out of our dataset (by filtering on the `published.print` field), 8 journals that met our inclusion criteria by being in the top 100 in impact factor in either political science or international relations, according to the Clarivate data we used. These are either online-only, or for some other reason don't have a `published.print` field in Crossref. Yet their articles have DOIs, and they should have been in our data.

The journals we are missing, and the number of papers for each, are:

journal	number of missing entries
All Azimuth: A Journal of Foreign Policy and Peace	126
Cuadernos Europeos de Deusto	266
Foro Internacional	454
Meridiano 47 - Journal of Global Studies	186
Monthly Review	6017
OASIS	157
Politics and Governance	559
Relaciones Internacionales	124

We are embarrassed by our inattentiveness to this, and only discovered it in a careful analysis of the data triggered by the question.

We do not think that this impacts our findings or the integrity of the paper. With the exception of "Politics and Governance," the other seven journals do not appear to publish much in the way of quantitative political science.

Note that the `.R` file with this analysis is in the `./rr_code/rr_crossref_date_field_choice_analysis.R` file, with comments.

As for the other questions, such as PDF conversion to text and more: We've updated the manuscript to make our process clearer.

Briefly: Optical character recognition was only used interactively on the command line, but the majority of the pipeline used `R` wrapped on the `pdftotext` command line utility, which uses Poppler, a well-known open source pdf wrangling library. The pdfs were mostly clean, and the conversions were high quality (again, based on randomly sampling them and looking at the outputs). URLs were converted like any other string. Of course, linebreaks and other issues meant that they were broken in some cases, and in those instances we simply painstakingly constructed (pre-ChatGPT) regular expressions to try to capture them (based on fragments of the URL), and stitched them back together. This is in the `dataverse_2_link_to_papers.R` file, for instance.

As for HTML – we in fact later did switch to html, and have just over 20,000 HTML files. Our reason for PDF in the first place was mostly convenience: We could use a GUI like Zotero to do large-scale downloading, which seems to have a better time accessing data from publishers than a scraper running in the browser.

(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.*

We've also expanded the discussion of this, including the steps we took to validate our methods and the numbers of papers that were involved in this step. We've also sharpened the language to be more explicit about what our methods can conclude. We had, in fact, neglected to include our results on what we've termed this 'precarious data' and so we appreciate the reviewer's diligence in calling our attention to these issues. Given that we are dealing with messy data and parsing strings with regular expressions, it is difficult to be certain that we caught everything – but by exporting multiple variations of results to Excel (i.e. with different combinations of strings, excluding Dataverse one way, including 'personal website' or 'http|www' another way) three times, and manually reviewing them, we felt we had a decent measure of the number of papers reporting precarious data of this sort.

(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. We further detailed in that section that we also searched for "preanalysis" and "pre-analysis". The `count_prereg_papers.R` file details the different terms we searched 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 and led to some changes throughout the manuscript. Many journals publish both types of papers, and it is conceptually fraught to dichotomize 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 for PSIR combined.

By way of background: We initially followed the typology of the Journal Citation Report 2020 (JCR), which are listed in Appendix Table 1. However, upon reflection these are somewhat

arbitrary, and after scrutinising the list more closely we find journals coded as solely PS or IR publish articles belonging to the other category. So we've dropped this distinction.

(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?*

We have a lot of thoughts and ideas for policies and technological affordances that could facilitate better practices, but they are still unrefined. We want to keep this a straightforward descriptive paper, and once we've done more research propose some recommendations for publishers and others in the scientific community to improve transparency in social science and beyond.