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

Bermond Scoggins* Matthew P. Robertson†

Latest version: 09 May, 2024

Abstract

The scientific method is predicated on transparency – yet the pace at which transparent research practices are being adopted by the scientific community is slow. The replication crisis in psychology showed that published findings employing statistical inference are threatened by undetected errors, data manipulation, and data falsification. To mitigate these problems and bolster research credibility, open data and preregistration practices have gained traction in the natural and social sciences. However, the extent of their adoption in different disciplines are unknown. We introduce computational procedures to identify the transparency of a research field using large-scale text analysis and machine learning classifiers. Using political science and international relations as an illustrative case, we examine 93,931 articles across the top 160 political science and international relations journals between 2010 and 2021. We find that approximately 21% of all statistical inference papers have open data and 5% of all experiments are preregistered. Despite this shortfall, the example of leading journals in the field shows that change is feasible and can be effected quickly.

*PhD Candidate, School of Politics and International Relations, Australian National University. Corresponding author. Email: bermond.scoggins@anu.edu.au.

†PhD Candidate, School of Politics and International Relations, Australian National University. Senior Research Associate, School of Social Sciences, University of Mannheim. Email: matthew.peter.robertson@uni-mannheim.de

1 Introduction

The Royal Society has as its motto the injunction *Nullius in verba*: "Take nobody's word for it." Yet a large portion of published studies in the social sciences demand of the reader exactly this.

Over the past several decades, open science advocates have called for the routinization of open science practices such as posting data and code upon a paper's publication and the preregistration of experiments (King 1995). Beginning principally in the psychological sciences, advocacy for these reforms rose in the 2010s due to large-scale replication failures of prominent psychological studies which highlighted the widespread presence of false positive findings (J. P. Simmons, Nelson, and Simonsohn 2011; Open Science Collaboration 2015).

Open science practices bolster the credibility of a field and its findings by allowing readers to evaluate the methods by which researchers reach their conclusions. While trust is the currency of every epistemic community, the demand for trust alone weakens credibility. If data and code are available, interested researchers can ensure a finding's results are computationally reproducible, robust to alternate model specifications, and error free. For experiments (i.e. randomised controlled trials), preregistration allows the reader to determine whether there was the selective exclusion of hypotheses, measurements, or statistical analyses that run counter to the author's favored hypotheses.

Concern for research transparency has become more salient over the past decade as scholars recognize that the accumulation of false positives can drive unsuccessful decision-making and interventions. This leads to inefficient resource allocation and weakens the credibility of a field. In fields like medicine, open science practices have been strongly advocated in recognition of the direct harm that false positives can cause (National Academies of Sciences and Medicine 2021; Baggerly and Coombes 2009). Leading journals in political science and international relations are increasingly mandating the provision of data and code, as well as encouraging the preregistration of experiments.

We distinguish the practice of making data and code available post-publication, thereby allowing researchers to determine what we hereafter refer to as the computational reproducibility of a paper's results, from replicibility – where new data is collected using an identical or conceptually similar design to the original paper (Nosek and Errington 2020; Obels et al. 2020). Different fields refer to these practices in ways that can be confusing. Political science, unlike psychology, conducts fewer experimental studies and what they refer to as a replication study aims at assessing computational reproducibility (see King 1995; Monroe 2018).

Political science and international relations appear to have taken open science practices seriously, with high-profile journals and academics endorsing initiatives like the Data Access and Research Transparency (DA-RT) statement. This has lead some scholars to believe that the open data problem has mostly been solved. Yet current assessments of the field's progress have been based on relatively small samples and time-intensive human coding procedures (Key 2016; Stockemer, Koehler, and Lentz 2018; Grossman and Pedahzur 2020).[1]

Our paper presents the largest-scale study of open science practices in political science and international relations thus far; it is also the first systematic study of the prevalence of preregistration in experiments in these fields. Our study spans the years 2010 to 2021 and includes population-level data, allowing us illustrate trends in specific journals. Documenting such trends is important given the key role played by journals in promulgating and enforcing transparent research norms.

We ask two questions: (1) What proportion of papers that rely on statistical inference make their data and code public? (2) What proportion of experimental studies were preregistered? We gather 93,931 published articles from the top 160 journals ranked by Clarivate's Journal Citation Reports (2020) and use machine learning classifiers to identify

[1]Key (2016) analyses 586 articles in six top political science and international relations journals – some of which have already adopted compulsory data availability policies – for 2014 and 2015. Stockemer, Koehler, and Lentz (2018) analyse data availability in the articles of three journals in 2015. Grossman and Pedahzur (2020, 1) analyze 92 articles published in the Fall 2019 issues of six journals and argue that the field is now approaching a "replicability utopia".

either statistical inference or experimental papers.[2] We identify which had open data and preregistration using public application programming interfaces (API), text analysis, and web scraping.

1.1 The state of open political science practices

Since the onset of the replication crisis, how much of the literature dependent on data and statistical inference still relies solely on reader trust? Extant research on the prevalence of open data practices in political science paints a sobering picture. Stockemer, Koehler, and Lentz's (2018) analysis of 145 quantitative studies published in three journals during 2015 found that only 55% provided original data and 56% provided code.[3] An earlier analysis, conducted on 494 quantitative articles in six leading political science journals between 2013 and 2014, found that full computational reproducibility materials (data and code) were available for only 58% of papers (Key 2016).[4]

Poor data availability affects many natural and social science disciplines (Culina et al. 2020; Errington et al. 2021). A random sample of 250 psychology papers published between 2014 and 2017 estimated that 14% of papers shared research materials, 2% provided original data, and 1% shared their code (Hardwicke, Thibault, et al. 2021). Preregistration was rare (3%). Similarly, even once data is shared, analytic reprodubility is not guaranteed (Hardwicke, Bohn, et al. 2021).

A tonic for many of these problems is straightforward: computational reproducibility materials for all quantitative studies and preregistration for experiments. Reproducibility materials and preregistration militates against questionable research practices (QRPs) that lead to false positives by constraining researcher degrees of freedom and ensuring that key

[2] A complete list of the journals can be found in the appendix.

[3] The three journals are *Electoral Studies*, *Party Politics*, *Journal of Elections, Public Opinion, and Parties* were analysed.

[4] The six journals analysed were *American Political Science Review*, *American Journal of Political Science*, *British Journal of Political Science*, *International Organization*, *Journal of Politics*, *Political Analysis*.

decisions made in the analysis process are transparent to peers.

In the behavioural sciences, false positives can arise from decisions that are rationalised as legitimate by authors: failing to report all dependent variables in a study, collecting more data after seeing whether the results were statistically significant, failing to report all conditions, stopping data collection after achieving the desired result, rounding down p-values, selectively reporting studies that 'worked', selectively excluding observations, and claiming an unexpected finding was predicted (or hypothesising after results are known). However, these practices obfuscate the uncertainty around a particular set of claims and mislead readers into being overconfident about a study's conclusions.

The use of QRPs appears to be widespread in many of the social sciences. Surveys of psychology and criminology researchers report they routinely do not report all dependent variables, collect more data after peeking at results, and selectively report statistically significant studies (John, Loewenstein, and Prelec 2012; Chin et al. 2021). Other methods of detecting publication bias, such as analysing sets of studies or literatures using a p-curve or z-curve, reveal extensive clustering of p-values (z-scores) just past $p < 0.05$ (Simonsohn, Nelson, and Simmons 2014; Bartoš and Schimmack 2020). Examples of these problems in the behavioural and social sciences range from the power posing literature (J. P. Simmons and Simonsohn 2017) to economic research using instrumental variables and difference-in-differences (Brodeur, Cook, and Heyes 2020).

In recognition of these problems, professional organisations in political science and international relations, including the American Political Science Association (APSA), have led efforts to increase the availability of data and code that accompany published papers. The DA-RT statement developed by the APSA council in 2014 involved a commitment by journal editor signatories to increase the availability of data "at the time of publication through a trusted digital repository", as well as require authors to "delineate clearly the analytic procedures upon which their published claims rely, and where possible to provide access to

all relevant analytic materials" (Statement 2015).

While there was an intramural debate about how DA-RT standards would affect qualitative work, given the heterogeneity of interview data and other forms of qualitative analysis, we bypass these arguments in this paper by focusing exclusively on papers relying on statistical inference.[5] It is relatively straightforward for researchers using statistical inference to release the very data and code that were necessary to produce the results in their papers. As Key (2016) notes, the internet has reduced the cost for journals to set up Dataverse repositories and made it easier for researchers to share their data and code. Rising usage of free statistical programming software, such as `R` and its desktop application RStudio, also reduces barriers to computational reproducibility.

The 27 journal editors who adopted the statement agreed to implement reforms by January 2016. Of the 16 DA-RT signatory journals in our dataset, two made no change in practice and a further four have data and code that is difficult to accurately estimate.[6]

1.2 The need for open data

1.2.1 Uncovering data errors and misinterpretation

Errors in data or the misreporting of p-values or test statistics invariably occur in research and can go undetected by an article's authors or peer reviewers. These problems, if addressed, may substantively alter an article's conclusions or produce null rather than positive results. Reporting errors in regression coefficients or test statistics occur frequently (Nuijten et al. 2016).

Access to the original data can help determine whether errors are trivial, and contribute to retraction efforts if they are not ("Retraction Notice" 2020). In some cases, access to the data allows for detailed concerns with a paper's analysis to be illustrated without the journal

[5]Summaries of these debates can be found in Lupia and Elman (2014) and on the Dialogue on DART website ("Perspectives on DA-RT," n.d.).

[6]We discuss these issues further in the results section.

believing a retraction is warranted (see Hilgard 2020, 2021).

1.2.2 Identifying model misspecification and facilitating extension

Researchers have tremendous flexibility in deciding how to collect data and which statistical models should be specified to analyse them. Andrew Gelman has termed this process the 'garden of the forking paths' (2014): some set of decisions might yield a positive result, while another set of equally justifiable decisions might lead to a null result. The replication crisis has shown that it is a mistake to view a single study or set of statistical analyses as a definitive answer to a given theory or claim — the scientific process should instead be iterative, exploratory, and cumulative (Tong 2019).

Open data can address the problem of model misspecification and uncertainty around modelling the data generating process (Neumayer and Plümper 2017). Since researchers cannot anticipate changes to methodological best practices, computational reproducibility materials allow scholars to make adjustments if best practices change.[7] Even if misspecification is not a problem, extending and building off of the original analyses – to run more theoretically motivated models, sensitivity analyses, or assess treatment heterogeneity – are net positives for science (Janz 2016).

1.2.3 Exposing data falsification

In the most egregious cases, open data allows researchers to investigate and expose data falsification. High-profile exposures of data falsification include the LaCour and Green (2014) case in political science, and the Shu et al. (2012) case in psychology (see Kristal et al. 2020; Nelson, Simonsohn, and Simmons 2021). Both rested on investigator access to the original data. While presumably data falsification is exceedingly rare, there is no way to know its extent given the general absence of computational reproducibility materials in the first place.

[7]For instance, Lenz and Sahn (2021) find that over 30% of observational studies published in the *American Journal of Political Science* rely on suppression effects to achieve statistical significance. Being able to determine the influence of suppression effects requires access to the original data.

1.3 The need for preregistration

1.3.1 Distinguishing confirmatory from exploratory analysis

Preregistration means that researchers specify their hypotheses, measurements, and analytic plans prior to running an experiment. This commits researchers to making theoretical predictions before they can view the data and be influenced by observing the outcomes (J. P. Simmons, Nelson, and Simonsohn 2011; J. Simmons, Nelson, and Simonsohn 2021). By temporally separating predictions from the data that tests their accuracy, there is much less flexibility for both post hoc theorising and alterations of statistical tests to fit the prediction.

Post hoc theorising, also known as hypothesising after the results are known (HARKing), is an example of circular logic — the researcher conducts many tests when exploring a dataset, the data reveals a relationship that can be made into a hypothesis, and that hypothesis is 'tested' on the data that generated it (Nosek et al. 2018). But the diagnosticity of a p-value is in part predicated on knowing how many tests were performed: when an exploratory finding is reported as a prediction, the normal methods employed to evaluate the validity of a hypothesis — such as whether the p-value is less than 0.05 (i.e. null hypothesis significance testing) — no longer hold. P-values in that case have unknown diagnositicity (Nosek et al. 2018). Thus, post hoc theorising and selective reporting greatly contribute to false positives.

1.3.2 Reducing the selective reporting of results

The selective reporting of statistical tests and results can occur for a variety of reasons. There are numerous legitimate ways of analyzing data, and this makes selective reporting seem justifiable. Danger arises when researchers convince themselves that the measures and tests lending evidence to their claims are the 'right' ones, while unjustifiably failing to report measures and tests that did not support the favored hypothesis.

Selectively reported experimental studies often result in overconfident theoretical claims and inflated effect sizes when compared to replications. The Open Science Collaboration

(2015) and Many Labs studies (2014, 2018) have shown that the effect sizes in highly powered replications are much smaller than those in the original studies. When reforms are implemented mandating preregistration, by research bodies or formats like registered reports, the number of null results reported rise (Kaplan and Irvin 2015; Scheel, Schijen, and Lakens 2021).

The primary purpose of preregistration is to provide journal reviewers and readers the ability to transparently evaluate predictions and the degree of flexibility researchers had to arrive at their conclusions (Lakens 2019; Claesen et al. 2019; Franco, Malhotra, and Simonovits 2014). It is up to the reader to determine whether preregistered studies followed their preregistration plan and adequately justified deviations – insufficiently detailed preregistration reports are an ongoing problem (Ofosu and Posner 2020).

The replication crisis has altered best practices and changed the habits of many researchers in the behavioural sciences. As we show below, preregistration is not yet the norm in political science and international relations. The conclusions from many studies relying on statistical inference, even some that that have been preregistered on a registry, remain exposed to the statistical pitfalls described above.

2 Methods

Our study design called for a comprehensive analysis of population-level data, yet our populations — (1) papers using data and statistics, and (2) original experiments — were embedded in a larger population of *all* political science and international relations publications in target journals. We downloaded all of the journals' papers from 2010 to 2021. Once we had these papers locally, we identified the data, statistical, and experimental papers through dictionary-based feature engineering and machine learning. We then used public APIs, web scraping, and text analysis to identify which of the studies had computational reproducibility materials. We outline this process below.

2.1 Phase one: gathering and classifying the papers

We used Clarivate's 2021 Journal Citation Report to identify target journals. We filtered for the top 100 journals in both political science and international relations, and combined the two lists for a total of 176 journals.[8]

With this list, we used the Crossref API to download all publication metadata. We were able to obtain records for 162 journals. This resulted in over 445,000 papers, which we then filtered on Crossref's `published.print` field for 2010 and onwards, resulting in 109,553 papers. We used the `published.print` field as it was the only reliable indicator of actual publication date, and the most complete.[9] As of mid-2023 we were able to obtain 93,931 of these articles in PDF and HTML formats, and we use this as the denominator in the study. We converted the PDFs to plaintext using the open source command line utility `pdftotext`, and we converted the HTML files to text using the `html2text` utility.

Identifying the papers that relied on data, statistical analysis, and experiments was an iterative process. In each case we read target papers and devised a dictionary of terms meant to uniquely identify others like them. We extensively revised these dictionaries to arrive at terms that seemed to maximally discriminate for target reports. The dictionaries eventually comprised 52, 180, and 133 strings, symbols, or regular expressions for the three categories respectively.

The dictionaries were then used with custom functions to create document feature matrices (DFM), where each paper is an observation, each column a dictionary term, and each cell a count of that term.[10] The DFM format made the papers amenable to large-scale analysis. In machine learning parlance, this process is known as feature engineering.

[8] As some journals publish both political science and international relations articles, the top 100 journals in each category overlapped.

[9] A more complete discussion of the choice of this field is found in the `./rr_code/rr_crossref_date_field_choice_analysis.R` file, which also shows the unintentional omission of seven journals.

[10] A custom function was preferable to existing text analysis libraries like `quanteda` because of our need to capture regular expressions and asterisks.

For the first research question – examining the presence of code and data in papers involving statistical inference – we hand-coded a total of 1,624 papers with boolean categories and identified 585 that relied on statistical inference. We defined statistical inference papers as any that involved mathematical modeling of data. This definition is meant to capture a simple idea: mathematical modeling requires computer instructions that perform functions on numbers. In the absence of computational reproducibility materials, these transformations cannot be exactly reproduced by readers. We also developed a dictionary of 35 terms for formal theory papers, because we wished to exclude papers that did not apply a model to real-world data.

For the second question — examining what proportion of experiments were preregistered — we hand-coded 518 papers with a single boolean category: whether the paper reported one or more original experiments. We defined this as any article containing an experiment where the researchers had control over treatment.

We then trained two machine learning models — the Support Vector Machine (SVM) and Naive Bayes (NB) binary classifiers — to arrive at estimates for the total number of statistical inference and experimental papers.[11] SVMs are a pattern recognition algorithms that give binary classifications to variables in high-dimensional feature space by finding the optimal separating boundary between labeled training data (James et al. 2021, 337–72; Cristianini and Shawe-Taylor 2000). The NB family of algorithms calculate the posterior probability of a given classified input based on the independent probability of all the values of its features; it then applies this trained algorithm to classify new inputs (Rhys 2020, 135–67).

We report the SVM model results both for their greater accuracy and due to our theoretical prior that the model would be more suitable for a high-dimensional classification

[11]As an additional robustness check to predict open data and statistical inference papers, we estimated a series of bivariate logistic regressions using the same DFMs. The predicted probability plots can be found in the appendix. These plots give a lower estimate than the machine learning models, though they are in the same broad range. We also attempted to use the Claude 3 Haiku model by Anthropic, but discontinued this experiment due to time and resource constraints, as detailed in the letter to reviewers in the replication materials.

problem. For the first research question, our SVM model achieved 92.35% accuracy for statistical papers. For the classifying experiments, the accuracy was 86.05%. In Appendix 1 we report the confusion matrices, hyperparemeter tuning data, and NB models.

The application of the SVM model to the full dataset of 93,931 publications leads to an estimate of 24,026 using statistical inference.

The identification of experimental papers proceeded slightly differently. Rather than beginning with the full corpus, we first filtered for only the papers that included the word "experiment" over five times (4,835). We then ran the SVM classifier on this subset. The resulting estimate was 2,552 papers reporting experiments.

2.2 Phase two: Identifying open data and preregistrations

We attempted to identify open data resources in seven ways.

1. Using the Harvard Dataverse API, we downloaded all datasets held by all journals in our corpus who maintained their own, named dataverse (n=20);
2. We queried the Dataverse for the titles of each of the 109,553 papers in our corpus and linked them to their most likely match with the aid of a custom fuzzy string matching algorithm. We validated these matches and manually established a string-similarity cut-off, setting aside the remainder;
3. We extracted from the full text of each paper in our corpus the link to its dataset on the Dataverse (1,142; note this had significant overlap with the results of the first and second queries);
4. We downloaded the metadata listing the contents of these datasets, to confirm firstly that they had data in them, and secondly that it did not consist of only pdf or doc files. In cases where a list of metadata was not available via the Dataverse API, we scraped the html of the dataset entry and searched for text confirming the presence of data files;
5. We used regular expressions to extract from the full text papers references to "replication

data," "replication materials," "supplementary files" and similar terms, then searched in the surrounding text for any corresponding URLs or mentions of author personal websites or other repositories[12]. We validated these results by exporting various combination of string matches with the above terms to Excel files, where we examined them in tabular format and validated their relevance. Given that replication and supplementary material stored on personal websites is not of the same order as material on the Dataverse and similar repositories, these results are recorded in our results under the rubric of 'precarious data';

6. We searched all of the full text papers for references to other repositories, including Figshare, Dryad, and Code Ocean, using regular expressions. Where found, these were recorded as containing replication data, the same as the Dataverse;
7. As additional validation for DA-RT signatory journals specifically, we downloaded the html file corresponding to each article and/or the html file hosting supplemental material (n=2,284), then extracted all code and data-related file extensions to establish their open data status.

We attempted to identify preregistration of experiments in the following ways:

1. We used regular expressions to extract from all of the experimental papers sentences that referred to "prereg" or "pre-reg", "preanalysis" or "pre-analysis", as well as any references to commonly used preregistration servers (OSF, EGAP, and AsPredicted), and then searched for the availability of the corresponding link to validate that the preregistration had taken place. Parts of this process — for instance, searching author names in the Experiments in Governance and Politics (EGAP) registry to look for the corresponding paper — involved time-consuming detective work;
2. We downloaded all EGAP preregistration metadata in JSON format from the Open Science Foundation Registry (https://osf.io/registries/discover), extracted from this

[12]Terms like "replication data" are used in political science to refer to computational reproducibility materials such as open data and code.

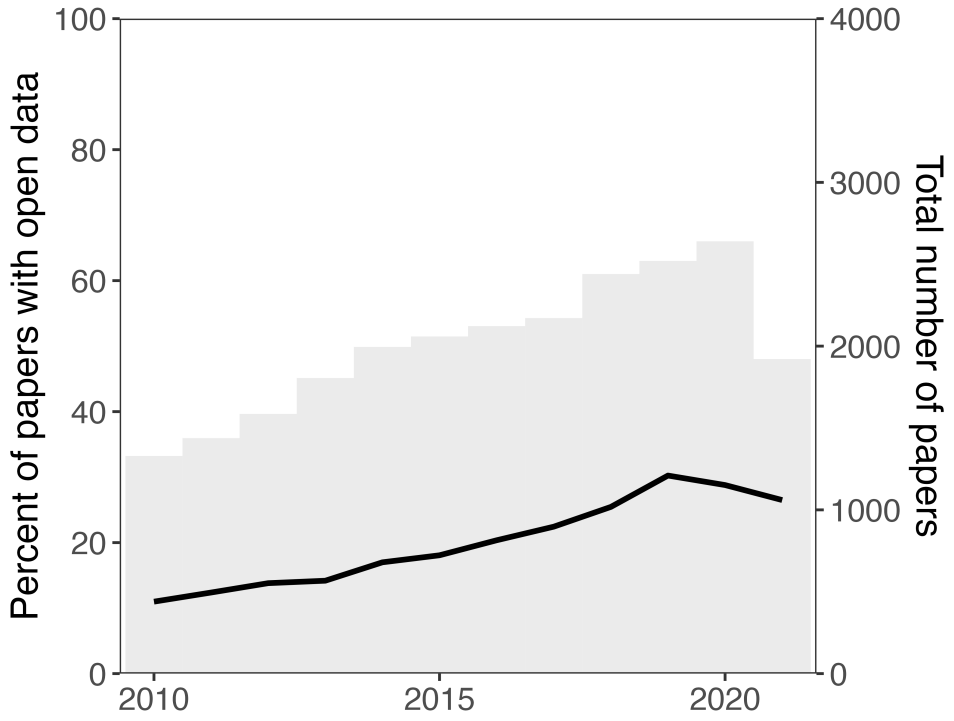

Figure 1: Open data in statistical inference papers by year

file all osf.io links and unique EGAP registry IDs, and used command line utilities to search for them through the corpus of all the papers.

We did not examine whether the published report conformed to the preregistration plan.

3 Results

Statistical inference papers are infrequently accompanied by the datasets or code that generated their findings. For the 12 year period under observation, we were able to match 21% of statistical inference articles to data repositories (overwhelmingly the Harvard Dataverse). Encouragingly, Figure 1 shows that the percentage of open data has increased between 2010 and 2021 – rising steadily from about 11% to 26% during this period.

The total number of statistical inference papers have gradually increased during the 12 year period. In 2010, we found 1,329 papers and 2,640 in 2020 – the last year with complete

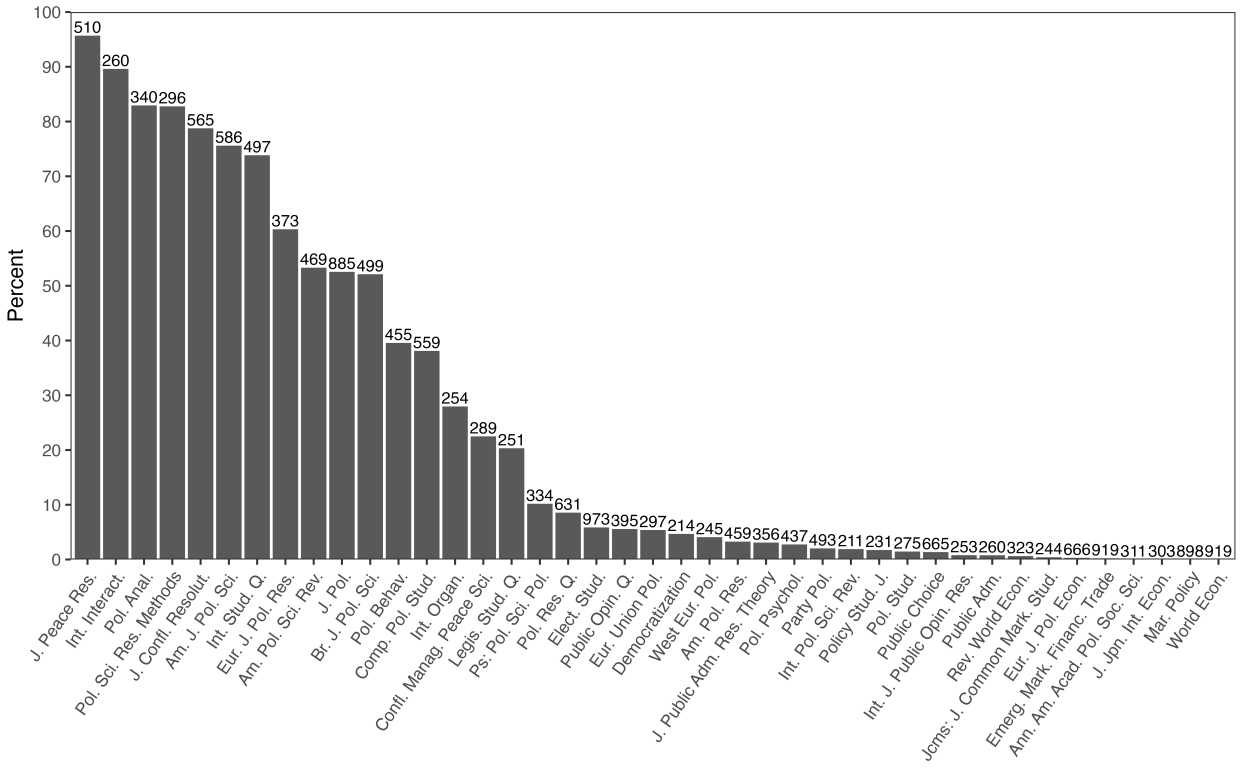

Figure 2: Open data in statistical inference papers by journal (with over 200 papers)

data. This supports King's (1990) observation that political science and international relations have long been disciplines increasingly concerned with quantitative methods.[13] While the percentage of papers with open data have increased, so too have the absolute number of statistical papers without it. There are simply more published papers making inferences based on hidden data.

There are significant differences in open data practices between journals. Figure 2 displays the percentage of statistical inference papers with open data in the 41 journals with over 200 such papers.[14] The number above each journal's bin represents the number of statistical inference papers detected by the support vector machine classifier. Of the 41

[13]Gary King illustrated that by 1988 almost half of publications in the American Political Science Review were quantitative.

[14]The cutoff was established to focus on journals who publish more quantitative papers and for ease of viewing – the graph with all 158 journals with at least one statistical inference paper is very large and is located in the appendix.

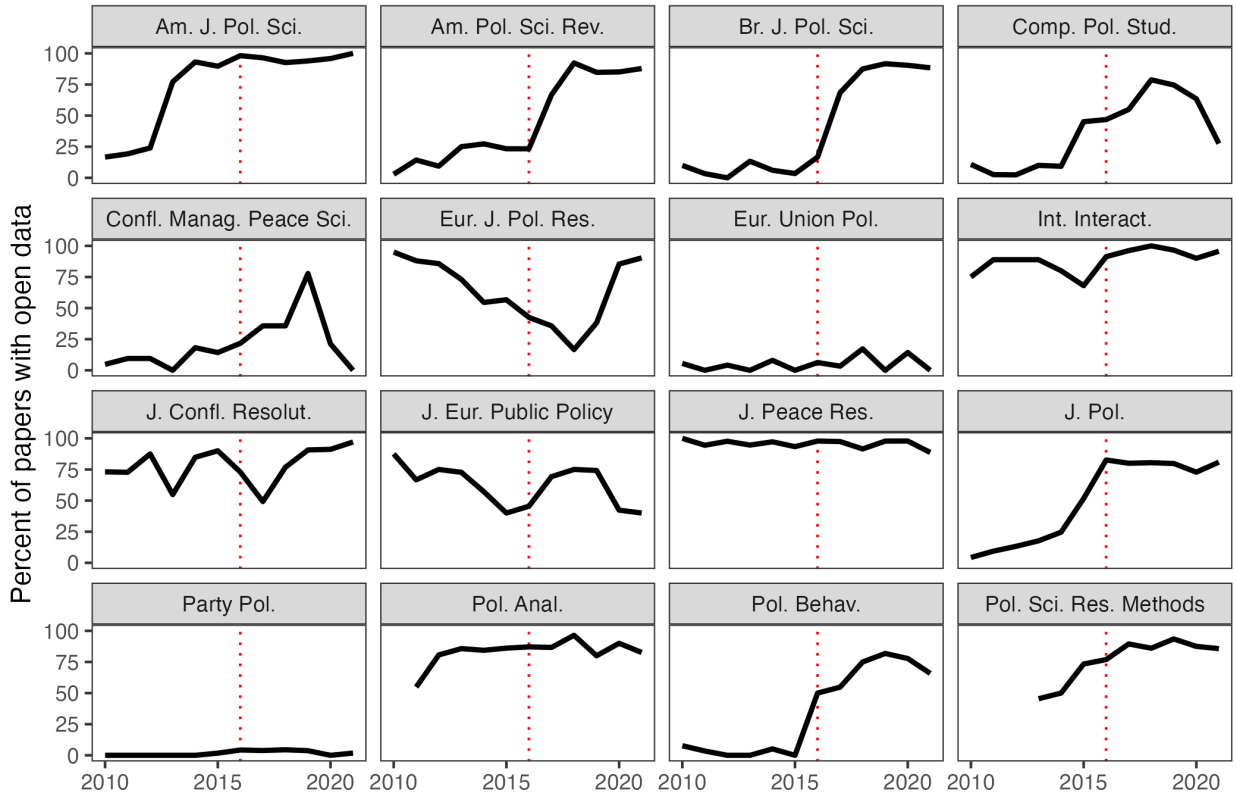

Figure 3: Open data in statistical inference papers by year published in 16 of the 27 journals signatory to the DA-RT statement

journals, 11 have over 50% open data, and 16 have over 20%.[15]

The effectiveness of the DA-RT statement on journal open data practices is illustrated in Figure 3, which displays the percentage of statistical inference papers with open data by year in each of the 16 DA-RT signatory journals we consider.[16]

[15]The journals with over 50% open data are the *American Journal of Political Science*, the *American Political Science Review*, the *British Journal of Political Science*, *European Journal of Political Research*, *International Interactions*, *International Studies Quarterly*, *Journal of Conflict Resolution*, *Journal of Peace Research*, *Journal of Politics*, *Political Analysis*, and *Political Science Research and Methods*. Those with over 20% open data include the aforementioned journals as well as *Comparative Political Studies*, *Conflict Management and Peace Science*, *International Organisation*, *Legislative Studies Quarterly*, and *Political Behaviour*.

[16]A total of 27 journals signed the DA-RT statement. The majority of these journals publish quantitative research (as can be seen in Figure 2). Note that there are actually 20 DA-RT signatory journals in our dataset, but four of them have an insignificant number of statistical inference publications and so we omit them from the analysis.

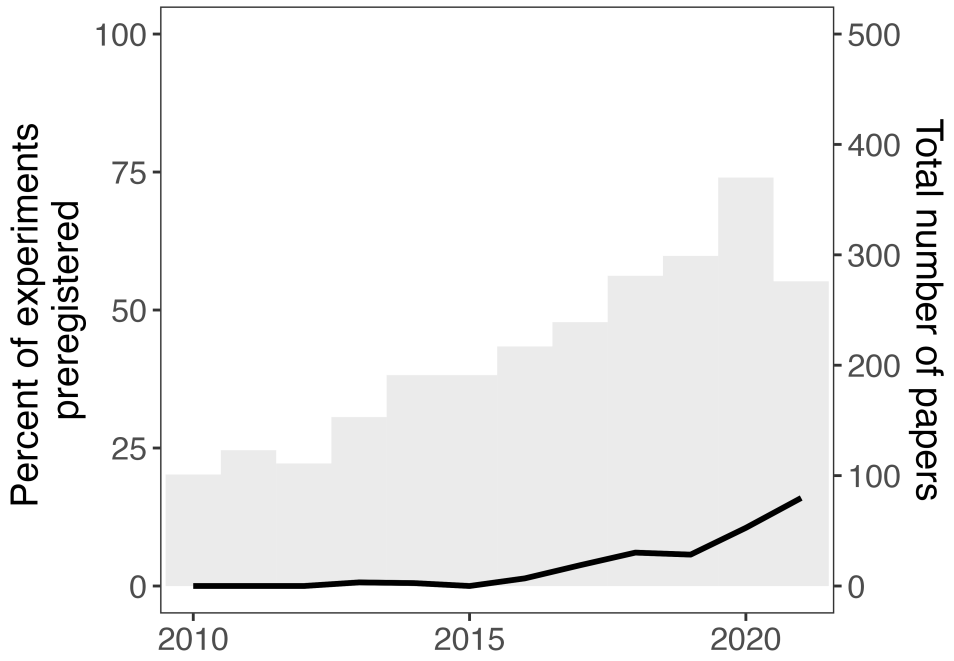

Figure 4: Preregistration in experiments by year

Four journals – *American Journal of Political Science*, *International Interactions*, *Political Analysis*, and *Political Science Research and Methods* – already made significant progress prior to the release of the initial DA-RT guidelines in 2014. Many of the remaining journals either made significant progress in 2016 or shortly thereafter.

One caveat is that is that 2 of the 16 journal signatories have consistently low levels of open data even after DA-RT reforms were agreed to commence on January 15, 2016. The extent of transparent practices in three other journals – *Journal of European Public Policy*, *European Journal for Political Research*, and *Conflict Management and Peace Science* – was more difficult to determine, given they did not use the Harvard Dataverse. Our attempt to estimate data and code availability for such journals, noted in point seven of phase two of the methods section, appears to produce unreliable and puzzling results.

The preregistration of experiments is rare in political science and international relations journals. Figure 4 shows that the first preregistered study in the dataset that we could

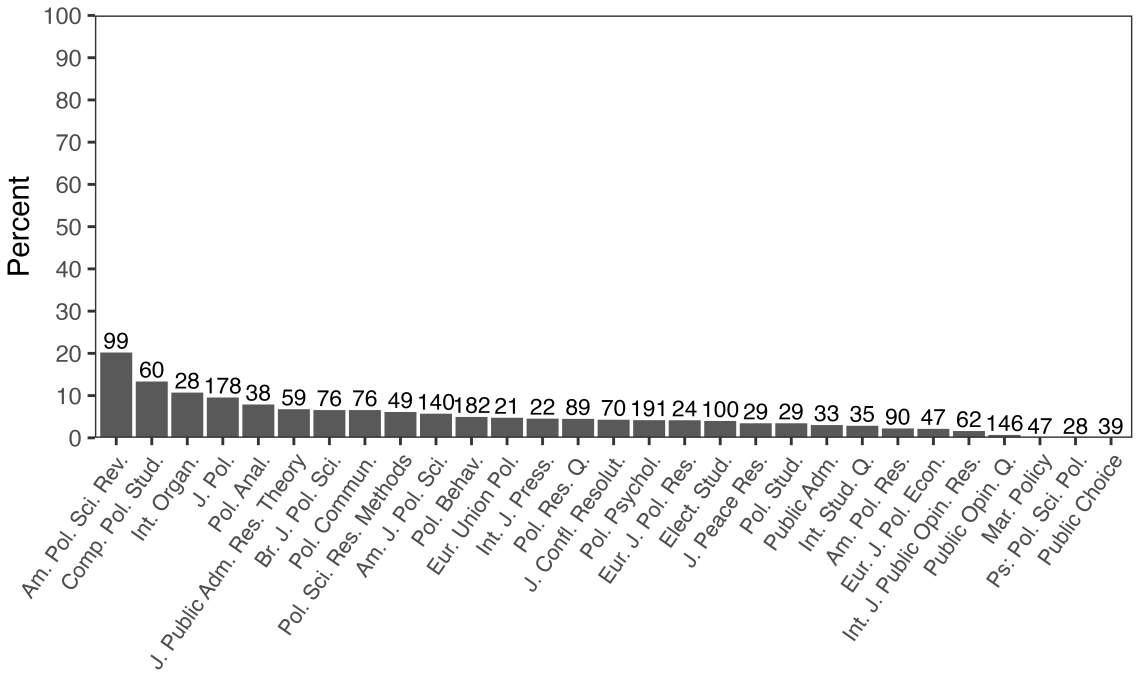

Figure 5: Preregistration in experiments by journal (with over 20 papers)

identify was in 2013, and that the rate of preregistration only began climbing in 2016. The proportion of experiments that were preregistered for the entire period is approximately 5%; the annual rate has slowly risen to 16% in 2021.

Figure 5 shows the percentage of experiments that were preregistered in the 29 journals with more than 20 experiments. Only the *American Political Science Review* exceeds 20%. Unlike with open data, when it comes to preregistration the differences between journals are small. Of the experiments published in *Political Psychology* and *Political Behaviour*, the two journals with the most experiments that bridge the gap between political science and psychology, only four and five percent respectively are preregistered.

Prior to the replication crisis at the beginning of the 2010s, there were no organized attempts at enforcing preregistration or using registered reports as a way of curbing researcher flexibility and its attendant QRPs. As psychology was among the first of the sciences to

reckon with its methodological issues, brought to light in part by such articles as Simmons, Nelson, and Simonsohn's (2011), it is logical that it took several years for these new practices to be adopted in contiguous disciplines like political science and international relations. But our data illustrate that significant improvements must be made in order for experiments in these fields to meet current methodological best practices.

4 Discussion

Scientists must carry out their work while simultaneously signalling and vouchsafing for its credibility. For the pioneers of the scientific method in 17th century Europe, this included an ensemble of rhetorical and social practices, including the enlistment of trusted witnesses to testify that experiments had in fact taken place as claimed – this is what Shapin refers to as the moral economy of science (Shapin 2018, 84, 107–8; 1995).

In the digital age, we argue that the credibility of social science must largely rest on computational reproducibility. The same goes for preregistration and experiments. Adhering to these practices ensures other social scientists can check and reproduce the findings, that the findings are valid, and also demonstrates a commitment to the norm of science as a shared enterprise.

The chief reason for depositing code and data is not for signalling: Open science practices provide the reader with an opportunity to transparently evaluate the evidence for a set of claims and scrutinise an article for any of the myriad problems that plague the use of data and statistical models. An interested reader could investigate an article's data and code for errors, determine whether results are robust to different model specifications, or, in rare cases, detect data falsification. For experiments, the published paper can be compared to the preregistration document to determine whether there were any unjustified deviations.

Our findings show that political science and international relations are not currently

living up to these best practices. For the approximately 25,000 statistical inference papers in the dataset, we could only identify approximately 21% that had a corresponding data respository. Despite improvement in most years, change has not been uniform across the discipline — most of the progress has been made by a handful of the highest impact factor journals. In 2020, for example, 16 out of the 52 journals with over 20 statistical inference papers had an open data percentage over 50% (see Figure A6) – 20 journals had an open data percentage over 20%. We could not locate data or code for two of the 16 DA-RT signatories in our dataset.

Universal open data is a collective action problem, and it is the responsibility of journals to foster and enforce these disciplinary norms. In the absence of that, individual researchers do not always share data, and requesting it can sometimes be mistaken as a gesture of challenge rather than collegiality. As Simonsohn (2013) notes, the modal response to his requests for original data was that the data was no longer available. We suspect that variation in open data practices between journals reflects differences in journal editors' views of its importance for research credibility.

The DA-RT initiative sparked spirited debate in the field about the provision of data and code — but the same cannot be said for preregistration. Experiments are rarely preregistered. Of the roughly 2,552 experiments in our dataset, 5% are preregistered. Given that the use of experiments only began to take off in 2014, as shown in Figure 3, the proportion of preregistered experiments in the literature is understandably low. Fortunately, the trend is positive. Two journals of 26 with more than 5 published experiments had a preregistration percentage of over 30% in 2020 (see Figure A7).

Identifying whether an experiment had a corresponding preregistration report was at times difficult. Numerous experiments made no mention of their preregistration report in the manuscript despite having one listed in a repository. Locating it was also difficult given changing manuscript titles and authors. Their omission in the manuscript is likely due to the

fact that many journal editors do not determine whether an experiment has a preregistration or pre-analysis plan or request their disclosure.[17]

The difficulty of matching an experiment with its preregistration report is far smaller than matching a manuscript to a concealed preregistration report. A unique and unanticipated problem we found were authors publishing a study where they omitted any reference to a preregistered experiment – ostensibly due to null findings. Byun, Kim, and Li (2021) use their survey data to make descriptive claims while failing to discuss the design or results of their experimental manipulation (Kim, Byun, and Li 2021). It is not clear whether their results failed to further their own argument or were possibly disconfirmatory. In either situation, readers are not permitted to transparently evaluate the strength of their claims.

Peer reviewers and readers of published works routinely examine whether a theory or explanation has appropriate evidence; whether the measurements are valid and reliable; whether the model has been appropriately specified. Here, we prompt referees and readers to also begin asking: (1) Are the computational reproducibility materials on the Harvard Dataverse or some other reliable repository? (2) Is the paper computationally reproducible based on those materials? (3) If an experiment, was it preregistered? (4) Does the analysis in the experimental paper follow the preregistration plan and are deviations from that plan justified?[18] We hope that evaluating scientific research in this manner will help move readers away from trusting research in the absence of open science practices to a more informed trust in their presence.

[17]Journals like the *Journal of Politics* require authors to disclose a preregistration report or justify why they did not preregister their experiment.

[18]For experiments, we acknowledge that these are by no means definitive criteria on which to judge the trustworthiness of a paper or finding. These practices should accompany efforts to build confidence in a finding through direct and conceptual replications.

5 Acknowledgements

Both authors acknowledge support from the Australian Government Research Training Program Scholarship, the ANU Library, Taylor & Francis, TeamViewer AG, and the Google Cloud Research Credits program (award GCP19980904).

6 References

Baggerly, Keith A., and Kevin R. Coombes. 2009. "Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology." *The Annals of Applied Statistics* 3 (4): 1309–34. https://doi.org/10.1214/09-AOAS291.

Bartoš, František, and Ulrich Schimmack. 2020. "Z-Curve 2.0: Estimating Replication Rates and Discovery Rates." https://psyarxiv.com/urgtn/download?format=pdf.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics." *Am. Econ. Rev.* 110 (11): 3634–60.

Byun, Joshua, DG Kim, and Sichen Li. 2021. "The Geopolitical Consequences of COVID-19: Assessing Hawkish Mass Opinion in China." *Political Science Quarterly* 136 (4): 641–65.

Chin, Jason M, Justin T Pickett, Simine Vazire, and Alex O Holcombe. 2021. "Questionable Research Practices and Open Science in Quantitative Criminology." *J. Quant. Criminol.*, August.

Claesen, Aline, Sara L B T Gomes, Francis Tuerlinckx, and Wolf Vanpaemel. 2019. "Preregistration: Comparing Dream to Reality."

Clarivate. 2020. "Journal Citation Reports." 2020.

Cristianini, Nello, and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.

Culina, Antica, Ilona van den Berg, Simon Evans, and Alfredo Sánchez-Tójar. 2020. "Low Availability of Code in Ecology: A Call for Urgent Action." *PLoS Biol.* 18 (7): e3000763.

Errington, Timothy M, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021. "Challenges for Assessing Replicability in Preclinical Cancer Biology." *Elife* 10 (December).

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Unlocking the File Drawer." *Science* 345 (6203): 1502–5.

Gelman, Andrew, and Eric Loken. 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: 460+.

Grossman, Jonathan, and Ami Pedahzur. 2020. "Can We Do Better? Replication and Online Appendices in Political Science." *Perspectives on Politics*, 1–6.

Hardwicke, Tom E, Manuel Bohn, Kyle MacDonald, Emily Hembacher, Michèle B Nuijten, Benjamin N Peloquin, Benjamin E DeMayo, Bria Long, Erica J Yoon, and Michael C Frank. 2021. "Analytic Reproducibility in Articles Receiving Open Data Badges at the Journal Psychological Science: An Observational Study." *Royal Society Open Science* 8 (1): 201494.

Hardwicke, Tom E, Robert T Thibault, Jessica E Kosie, Joshua D Wallach, Mallory C Kidwell, and John P A Ioannidis. 2021. "Estimating the Prevalence of Transparency and Reproducibility-Related Research Practices in Psychology (2014-2017)." *Perspect. Psychol. Sci.*, March.

Hilgard, Joseph. 2020. "Curious Features of Data in Zhang Et Al. (2019)." OSF.

———. 2021. "I Tried to Report Scientific Misconduct. How Did It Go?" 2021. http://crystalprisonzone.blogspot.com/2021/01/i-tried-to-report-scientific-misconduct.html.

James, Gareth Michael, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An*

Introduction to Statistical Learning: With Applications in R. Springer Nature.

Janz, Nicole. 2016. "Bringing the Gold Standard into the Classroom: Replication in University Teaching." *Int Stud Perspect* 17 (4): 392–407.

John, Leslie K, George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychol. Sci.* 23 (5): 524–32.

Kaplan, Robert M, and Veronica L Irvin. 2015. "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time." *PLoS One* 10 (8).

Key, Ellen M. 2016. "How Are We Doing? Data Access and Replication in Political Science." *PS Polit. Sci. Polit.* 49 (02): 268–72.

Kim, D. G., Joshua Byun, and Sichen Li. 2021. "Foreign Policy Revisionism in the Era of COVID-19: Theory and Evidence from Chinese Public Opinion." https://osf.io/r9dn7.

King, Gary. 1990. "On Political Methodology." *Polit. Anal.* 2: 1–29.

———. 1995. "Replication, Replication." *PS: Political Science & Politics* 28 (3): 444–52.

Klein, Richard A, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahnıék, Michael J Bernstein, Konrad Bocian, et al. 2014. "Investigating Variation in Replicability: A 'Many Labs' Replication Project." *Soc. Psychol.* 45 (3): 142–52.

Klein, Richard A, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams, Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–90.

Kristal, Ariella S, Ashley V Whillans, Max H Bazerman, Francesca Gino, Lisa L Shu, Nina Mazar, and Dan Ariely. 2020. "Signing at the Beginning Versus at the End Does Not Decrease Dishonesty." *Proc. Natl. Acad. Sci. U. S. A.* 117 (13): 7103–7.

LaCour, Michael J, and Donald P Green. 2014. "Political Science. When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality." *Science* 346 (6215): 1366–69.

Lakens, Daniel. 2019. "The Value of Preregistration for Psychological Science: A Conceptual Analysis." *PsyArXiv*.

Lenz, Gabriel S, and Alexander Sahn. 2021. "Achieving Statistical Significance with Control Variables and Without Transparency." *Polit. Anal.* 29 (3): 356–69.

Lupia, Arthur, and Colin Elman. 2014. "Openness in Political Science: Data Access and Research Transparency: Introduction." *PS Polit. Sci. Polit.* 47 (1): 19–42.

Monroe, Kristen Renwick. 2018. "The Rush to Transparency: DA-RT and the Potential Dangers for Qualitative Research." *Perspectives on Politics* 16 (1): 141–48.

National Academies of Sciences, Engineering, and Medicine. 2021. *Developing a Toolkit for Fostering Open Science Practices: Proceedings of a Workshop*. Edited by Thomas Arrison, Jennifer Saunders, and Emi Kameyama. Washington, DC: The National Academies Press. https://www.nap.edu/catalog/26308/developing-a-toolkit-for-fostering-open-science-practices-proceedings-of.

Nelson, Leif D, Uri Simonsohn, and Joseph P Simmons. 2021. "[98] Evidence of Fraud in an Influential Field Experiment about Dishonesty." http://datacolada.org/98.

Neumayer, Eric, and Thomas Plümper. 2017. *Robustness Tests for Quantitative Research*. Cambridge University Press.

Nosek, Brian A, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. 2018. "The

Preregistration Revolution." *Proc. Natl. Acad. Sci. U. S. A.* 115 (11): 2600–2606.
Nosek, Brian A, and Timothy M Errington. 2020. "What Is Replication?" *PLoS Biology* 18 (3): e3000691.
Nuijten, Michèle B, Chris H J Hartgerink, Marcel A L M van Assen, Sacha Epskamp, and Jelte M Wicherts. 2016. "The Prevalence of Statistical Reporting Errors in Psychology (1985–2013)." *Behav. Res. Methods* 48 (4): 1205–26.
Obels, Pepijn, Daniel Lakens, Nicholas A Coles, Jaroslav Gottfried, and Seth A Green. 2020. "Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology." *Advances in Methods and Practices in Psychological Science* 3 (2): 229–37.
Ofosu, George K, and Daniel N Posner. 2020. "Pre-Analysis Plans: An Early Stocktaking." *Perspectives on Politics*, 1–17.
Open Science Collaboration. 2015. "PSYCHOLOGY. Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.
"Perspectives on DA-RT." n.d. https://dialogueondart.org/perspectives-on-da-rt/.
"Retraction Notice." 2020. *Youth Soc.* 52 (2): 308–8.
Rhys, Hefin. 2020. *Machine Learning with r, the Tidyverse, and Mlr*. Simon; Schuster.
Scheel, Anne M, Mitchell R M J Schijen, and Daniël Lakens. 2021. "An Excess of Positive Results: Comparing the Standard Psychology Literature with Registered Reports." *Advances in Methods and Practices in Psychological Science* 4 (2).
Shapin, Steven. 1995. "Trust, Honesty, and the Authority of Science." In *Society's Choices: Social and Ethical Decision Making in Biomedicine*, edited by Ruth Ellen Bulger, Elizabeth Meyer Bobby, and Harvey V Fineberg, 388–408. National Academy Press.
———. 2018. *The Scientific Revolution*. University of Chicago Press.
Shu, Lisa L, Nina Mazar, Francesca Gino, Dan Ariely, and Max H Bazerman. 2012. "Signing at the Beginning Makes Ethics Salient and Decreases Dishonest Self-Reports in Comparison to Signing at the End." *Proceedings of the National Academy of Sciences*.
Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychol. Sci.* 22 (11): 1359–66.
Simmons, Joseph P, and Uri Simonsohn. 2017. "Power Posing: P-Curving the Evidence." *Psychol. Sci.* 28 (5): 687–93.
Simmons, Joseph, Leif Nelson, and Uri Simonsohn. 2021. "Pre-registration: Why and How." *J. Consum. Psychol.* 31 (1): 151–62.
Simonsohn, Uri. 2013. "Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone." *Psychol. Sci.* 24 (10): 1875–88.
Simonsohn, Uri, Leif D Nelson, and Joseph P Simmons. 2014. "P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results." *Perspect. Psychol. Sci.* 9 (6): 666–81.
Statement, Journal Editors' Transparency. 2015. "Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors." *Political Science Research and Methods* 3 (3): 421–21.
Stockemer, Daniel, Sebastian Koehler, and Tobias Lentz. 2018. "Data Access, Transparency, and Replication: New Insights from the Political Behavior Literature." *PS Polit. Sci. Polit.* 51 (4): 799–803.
Tong, Christopher. 2019. "Statistical Inference Enables Bad Science; Statistical Thinking

Enables Good Science." *Am. Stat.* 73 (sup1): 246–61.