Advance Access publication March 24, 2015

Political Analysis (2015) 23:306–312
doi:10.1093/pan/mpv006

Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results

Annie Franco
Department of Political Science, Stanford University, 616 Serra St. Stanford, CA 94305, USA
e-mail: abfranco@stanford.edu

Neil Malhotra
Graduate School of Business, Stanford University, 655 Knight Way, Stanford, CA 94305, USA
e-mail: neilm@stanford.edu
(corresponding author)

Gabor Simonovits
Department of Political Science, Stanford University, 616 Serra St.Stanford, CA 94305, USA
e-mail: gabor.simonovits@gmail.com

Edited by R. Michael Alvarez

The accuracy of published findings is compromised when researchers fail to report and adjust for multiple testing. Preregistration of studies and the requirement of preanalysis plans for publication are two proposed solutions to combat this problem. Some have raised concerns that such changes in research practice may hinder inductive learning. However, without knowing the extent of underreporting, it is difficult to assess the costs and benefits of institutional reforms. This paper examines published survey experiments conducted as part of the Time-sharing Experiments in the Social Sciences program, where the questionnaires are made publicly available, allowing us to compare planned design features against what is reported in published research. We find that: (1) 30% of papers report fewer experimental conditions in the published paper than in the questionnaire; (2) roughly 60% of papers report fewer outcome variables than what are listed in the questionnaire; and (3) about 80% of papers fail to report all experimental conditions and outcomes. These findings suggest that published statistical tests understate the probability of type I errors.

1 Introduction

Scientific research depends on the accurate reporting of empirical results. When research articles do not report the full set of estimated specifications—either because researchers choose not to include them when they submit papers to journals or because editors and reviewers require authors to "streamline" their manuscripts—the validity of published findings is compromised in two ways. First, such practices will overstate the precision of estimates because they conceal the uncertainty associated with testing multiple hypotheses (Humphreys, Raul Sanchez, and van der Windt 2013). Second, selective presentation of specifications allows researchers to "present anything as significant" (Simmons, Nelson, and Simonsohn 2011). If outcome reporting is biased in favor of statistically significant results, large effect sizes will be overrepresented in the published literature and the accumulation of scientific knowledge will be impeded (Gelman and Tuerlinckx 2000; Gelman, Hill, and Yajima 2012).

Recently, several scholars have proposed institutional changes to address issues related to research transparency such as the requirement that authors preregister studies in advance of their execution and file extensive preanalysis plans (e.g., Gerber and Malhotra 2008; Casey, Glennerster, and

Author's note: Supplementary materials for this article are available on the *Political Analysis* Web site. Replication data are available on the Dataverse site for this article, http://dx.doi.org/10.7910/DVN/28766.

© The Author 2015. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mpv006 Published online by Cambridge University Press

Miguel 2012; Asendorpf et al. 2013; Franco, Malhotra, and Simonovits 2014; Miguel et al. 2014).[1] Despite some obvious benefits, preanalysis plans have been criticized as potentially stifling valuable inductive reasoning and learning from data (Anderson 2013; Laitin 2013). Theoretically, this should not be a concern since preanalysis plans do not have to be binding: authors can report both the results of tests specified in the plan and inductive findings from exploratory analyses. The advantage is that the reader is aware of what was deductively predicted as well as the total number of hypotheses tested. However, in practice, if reviewers and editors harshly judge papers that do not confirm *a priori* predictions, then inductive discovery may indeed be stifled.

These debates have currently taken place in a vacuum with almost no data on the prevalence of underreporting. All changes in scholarly practice (via either shifting norms or institutional reforms) have costs and benefits. A clear benefit of preanalysis plans is the curbing of research practices that allow investigators to produce statistically significant results through data mining. But the magnitude of this benefit—compared with the potential costs of stifling induction—is unknown. One reason for this lack of evidence is that it is difficult to determine what researchers planned to do from what is actually reported in a publication.

The goal of this letter is to provide empirical evidence on the prevalence of research practices associated with specification search. We do so by leveraging Time-sharing Experiments in the Social Sciences (TESS), a competitive, NSF-sponsored program that runs survey-based experiments on representative samples of the U.S. adult population at no cost to researchers. More details about TESS—as well as a discussion of whether TESS studies are representative of political science research in general—can be found in online Appendix 1.

A key feature of TESS studies is that the questionnaire of each survey experiment is made publicly available on the program's website 1 year after it has been conducted. These questionnaires are written *before* the data are collected and analyzed. Although TESS proposals are not publicly available due to confidentiality concerns, the questionnaires provide many clues as to what the researchers intended to do with the data. Another key feature of TESS is that researchers face strict caps on the number of respondent-questions (i.e., the number of questions multiplied by the number of respondents asked each question) as part of the grant. Due to these constraints, it is likely that the costly inclusion of an item in the questionnaire reflects an important theoretical expectation outlined in the proposal.

We compare the experimental conditions and outcome variables found in questionnaires to those reported in the respective publications to gauge the extent of underreporting.[2] Our analysis reveals underreporting of both experimental conditions and outcome variables.[3] Before delving into the analysis, we note that the purpose of this paper is not to criticize particular authors of papers, because (1) journal articles reflect not only the behaviors of authors but also requests made by editors and reviewers and (2) scholars operate within the prevailing norms of the discipline. Accordingly, we have not identified any specific studies in this manuscript. Indeed, one of the authors of this letter wrote papers in the data set that are characterized by underreporting.

[1] Recent examples of social science research using preanalysis plans include King et al. (2007), Casey et al. (2012), and Monogan (2013). After a new initiative, the journal *Cortex* became the first psychology journal to review papers solely based on research design and prior to data collection (see Chambers 2013). Since 2013, several other journals in psychology have adopted stricter reporting requirements (e.g., *Psychological Science*, *Experimental Psychology*, and *Social Psychology*).

[2] A previous version of this letter also compared the number of other covariates (such as pretreatment variables) measured in the surveys to the number reported in the published articles. However, our findings with respect to underreporting of these additional variables are less precise. It is difficult to discern whether authors intended to use such items to estimate heterogeneous treatment effects, for covariate adjustment, or for some other purpose. Consequently, questionnaire items falling into this "other items" category are coded with some degree of measurement error. A detailed description and the full set of results can be found in online Appendix 4.

[3] Additional complications, not explored in this paper, include the ability of researchers to run multiple experiments and only report the ones which yield significant results, stopping rules where researchers collect observations up until the point where results exceed arbitrary significance thresholds, and subsetting the data in arbitrary ways that make it likely to reject the null hypothesis. We do not explore these complications because we cannot learn anything about these practices from the survey instrument itself.

https://doi.org/10.1093/pan/mpv006 Published online by Cambridge University Press

Table 1 Distribution of journals of published TESS studies (2002–2012)

Journal	*Number of articles*
American Journal of Political Science	4
American Political Science Review	1
American Politics Research	1
British Journal of Political Science	1
Comparative Politics	1
Election Law Journal	1
Foreign Policy Analysis	2
International Organization	2
International Studies Quarterly	1
Journal of Experimental Political Science	1
Journal of Politics	7
Legislative Studies Quarterly	1
PS: Political Science and Politics	1
Political Analysis	1
Political Behavior	7
Political Psychology	4
Political Research Quarterly	2
Political Science Quarterly	1
Presidential Studies Quarterly	1
Public Choice	1
Public Opinion Quarterly	10
Quarterly Journal of Political Science	1
State Politics & Policy Quarterly	1
Total	53

2 Data and Measurement

2.1 *Sample*

The initial sample consisted of the entire online archive of TESS studies as of January 1, 2014, or 249 studies conducted between 2002 and 2012. The analysis reported below is restricted to the 53 studies whose results have been published in peer-reviewed political science journals (see Table 1 for the full list of journals). Additional details on the identification of published studies can be found in online Appendix 2. For each of the published studies we also downloaded the corresponding questionnaire used by GfK/Knowledge Networks to execute the survey (publicly available at http://www.tessexperiments.org/previousstudies.html). We used these questionnaires to extract the full list of measured variables, the logic of the experimental design, and the role of measured variables in the statistical analysis (replication data available at Franco, Malhotra, and Simonovits 2015).[4]

[4]TESS requires investigators to summarize and disseminate the results of their studies on the program website. Online summaries are available for most TESS studies and describe their experimental design, hypotheses, and main findings. However, these summaries are not especially useful for measuring underreporting in published research. First, they are completed by TESS PIs *after* the data are received and analyzed, making it difficult to distinguish what is pre-hoc from what is post-hoc. Second, the summaries often lack sufficient detail in the description of the specific variables used in the analysis. For instance, the online abstracts describe categories of outcome variables without specifying which survey items comprise the category. Finally, of the 52 available online summaries, 10 do not fully report the experimental treatments conducted in their corresponding study. Given that experimental manipulations are almost certainly reflective of important theoretical expectations and the abstracts are not always transparent about this element of the research design, we are not confident that these summaries would be fully informative of researchers' hypothesized outcomes.

https://doi.org/10.1093/pan/mpv006 Published online by Cambridge University Press

2.2 *Design Features*

We compare two aspects of experimental design as planned in the questionnaire and reported in the published papers: (1) experimental conditions and (2) outcome variables.[5] We classified items appearing in each questionnaire into one of these two categories and checked whether they appeared in the corresponding published article.[6] An example of how the coding procedure was implemented for an exemplary article can be found in online Appendix 3.

2.2.1 Experimental conditions

Experimental design is fundamentally characterized by the random assignment of subjects to experimental groups. Scholars may have *a priori* hypotheses of how survey responses might differ across these experimental conditions. They might also selectively report conditions that exhibit statistically significant differences from one another (or support their hypotheses) and exclude conditions that do not exhibit significant differences (or run against their expectations).

We gauged the extent to which researchers fully report the experimental manipulations executed in their studies. In the case of experiments that manipulated only one variable, we counted the number of conditions in the survey and the number reported in the corresponding article. In the case of experiments with factorial designs (i.e., those manipulating multiple variables), we counted the number of experimental treatment arms in the questionnaire and the number reported in the published paper.[7] This coding rule (1) provides an accurate accounting of the choices made by TESS researchers at both the design and analysis stages, (2) facilitates a direct comparison of the two, and (3) is not skewed by the large number of treatment groups in experiments with factorial designs.

2.2.2 Outcome variables

In survey experiments, researchers expect that the treatment stimuli will affect responses to subsequent questionnaire items. We coded items as "outcomes" if they were asked after the experimental treatment and could plausibly be affected by the treatments.[8] Researchers also often include manipulation checks, or post-treatment variables designed to assess whether the treatment is working as intended, but might be tempted to not report them if the results are not as hypothesized. Although one can argue that manipulation checks are less important and therefore acceptable to not report, a failed manipulation check is usually evidence that the treatment is not manipulating what it is theoretically supposed to. Moreover, in the absence of correcting for multiple comparisons, the failure to report manipulation checks will bias the reported *p*-values of the other outcome variables downward.

[5]We coded a design feature as being reported if it was mentioned in the text of the published article. A subtle form of underreporting is to describe some specifications in great detail and mention others only briefly. However, to avoid taking a stance about the boundaries between the reporting and "framing" of results, we used this relatively liberal coding rule.

[6]Each questionnaire-article pair was coded by one author and double-checked by one other author. The agreement rate for the two design features explored here was 94.3% (two discrepancies were with respect to whether a post-treatment variable should be counted as an outcome variable; one other discrepancy was with respect to whether an experiment employed a factorial design). In case of disagreement, the three authors collectively decided on the coding.

[7]A study with a control, treatment, and placebo group, for instance, would be coded as having three experimental conditions. In the case of factorial designs, we coded each factor as a separate treatment arm. A study that has a control group and interacts two factors with four levels each (a $4 \times 4 + 1$ design, with 17 experimental conditions) would also be coded as having three experimental conditions. We applied this coding scheme to publications as well. If an article failed to report the results from one or more levels of a factor, it would be coded as not reporting the corresponding treatment arm.

[8]In some cases we chose not to count items asked after the experiments as outcomes if it seemed implausible that they were meant to be used as dependent variables (e.g., number of children). It is possible that researchers placed these items post-treatment to guard against priming responses to the experimental stimuli.

https://doi.org/10.1093/pan/mpv006 Published online by Cambridge University Press

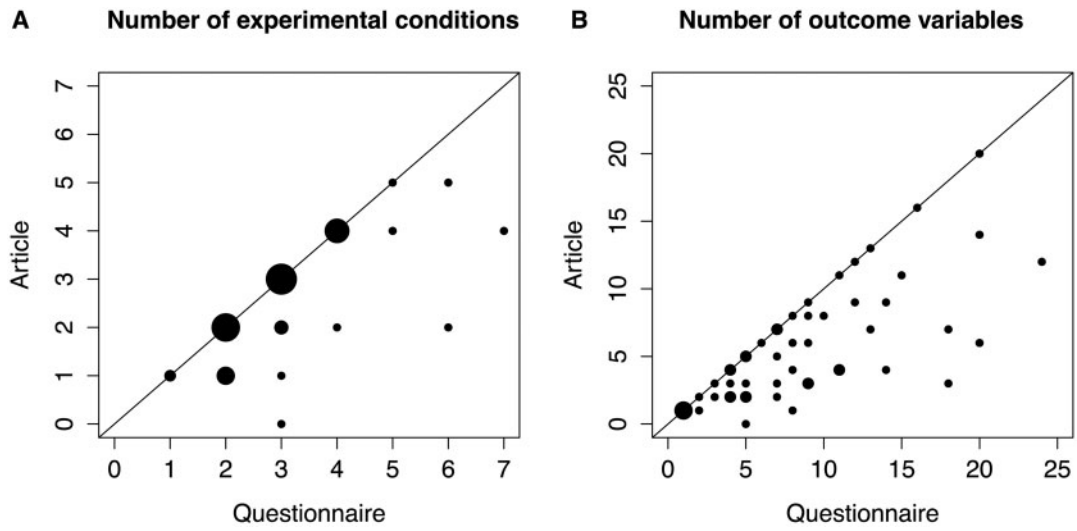


Fig. 1 Comparing design features in questionnaires and published results in TESS studies.
Note. Point size in each panel is proportional to the number of studies with a particular questionnaire-article value pair.

2.2.3 Limitations

Two caveats are in order. First, researchers might leverage a given TESS study to implement several experiments that they intend to publish in different papers. In cases where articles explicitly mentioned other papers reporting on omitted conditions, experiments, or outcome variables, we coded experimental conditions as being reported when we were able to find those papers. Very few articles mentioned such omitted conditions. However, it is possible that some of the experiments or conditions we coded as unreported *will* be reported in future articles. Further, some of these unpublished secondary papers could be due to reviewer rejections.

Second, our research design only measures underreporting, not *selective* underreporting. In other words, we do not analyze whether the survey items the authors reported are more likely to be associated with statistically significant results or results that exhibit larger effect sizes. Given the incentives to publish statistically and substantively significant results, we assume that this constitutes much of underreporting but cannot be completely sure without analyzing each data set individually. Regardless of the underlying mechanisms or whether it is selective, underreporting implies that the *p*-values presented in research from a given experiment are, overall, too low.

3 Results

We illustrate the descriptive results in Fig. 1. For each design feature, we plot the number of items listed in the questionnaires against the number mentioned in the respective published papers. Observations below the 45-degree line are indicative of underreporting; observations closer to the lower-right part of the figure indicate more severe instances of underreporting. As shown in Fig. 1, almost 30% of papers report fewer experimental conditions in the published paper than in the questionnaire (panel A). Overall, 60% of papers report fewer outcome variables than what are listed in the questionnaire (panel B). Examining the joint distribution of underreporting across design features reveals that barely a fifth of studies in our sample reported all experimental conditions *and* outcome variables.

Table 2 reports the mean number of design features for questionnaires (abbreviated as Q), articles (A), and the difference (Q–A). We also report 95% confidence intervals for both the means and mean differences. To avoid making parametric assumptions about the distribution of underreporting as a random variable, we employ non-parametric bootstrapping[9] to calculate

[9]Parametric estimation of standard errors and confidence intervals is ill advised given our small sample size and, more importantly, because underreporting is not distributed normally (it is non-negative by definition).

https://doi.org/10.1093/pan/mpv006 Published online by Cambridge University Press

Table 2 Underreporting of design features in published TESS studies

Design feature		*Mean*	*SE*	*95% CI*
Experimental conditions	*Q*	3.1	0.2	[2.8,3.5]
	A	2.7	0.2	[2.4,3.0]
	Q–A	0.5	0.1	[0.2,0.7]
Outcome variables	*Q*	8.6	0.8	[7.1,10.2]
	A	5.5	0.6	[4.4,6.7]
	Q–A	3.1	0.5	[2.1,4.1]

Note. *Q* refers to the mean number of design features included in the survey questionnaire. *A* refers to the mean number of design features reported in the published paper. *Q*−*A* represents the degree of underreporting. Standard errors and confidence intervals calculated by drawing 100,000 bootstrap replications.

standard errors and confidence intervals for each quantity of interest.[10] As shown in Table 2, the average questionnaire mentioned 3.1 experimental conditions while the average paper only mentioned 2.7 conditions (95% confidence interval around the mean difference is 0.2–0.7). Questionnaires included 8.6 outcome variables while the articles only reported 5.5 of these outcomes (95% confidence interval around the mean difference is 2.1–4.1).

4 Conclusion

We compare the questionnaires of a known population of published survey experiments to their published results and find that published papers diverge from research protocols, with underreporting of outcome variables and experimental manipulations. These practices lead to underestimation of type I errors. If this underreporting is selective, then published effect sizes are likely to be overestimated.

Preregistration of research protocols may help address selective reporting in studies by making them observable to the academic community. However, such proposals are unlikely to attenuate outcome and treatment reporting biases. This is consistent with research in the biomedical sciences that shows that published randomized trials continue to exclude negative outcomes or statistically insignificant results despite clinical registry requirements (Chan et al. 2004; Chan and Altman 2005; Rising, Bacchetti, and Bero 2008; Moreno et al. 2009). Although preregistration and replication archives may help identify these practices, the evidence presented here suggests that preanalysis plan requirements may be the most effective mechanism for curtailing such research practices. At the same time, if the problem is largely due to editors' and reviewers' requests to "streamline" manuscripts, then preanalysis plans may have limited value. One potential solution is to encourage authors to report the full set of analyses in online appendices and adjust standard errors for multiple testing, even if only a partial set of analyses are reported in the main text for framing purposes.

In closing, we want to emphasize that the pattern we have documented here should not be interpreted as fraud but rather as a research practice, which is driven by norms prevalent in the discipline. Our main point is that these norms and practices compromise learning from data and the accumulation of knowledge. Political science faces a collective action problem: individual researchers will not deviate from standard research practices unilaterally and incur professional costs until the field changes as a whole so that researchers are required (or at least incentivized) to fully report scholarly findings. Along with these changes in practice may come other benefits, such as the increased likelihood of publishing null findings and the evaluation of research projects based on the quality of design (as opposed to exceeding arbitrary significance levels).

[10]We take 100,000 bootstrap replications and calculate standard errors and confidence intervals from the respective bootstrap distributions of mean underreporting of each design feature. This procedure is robust to the correlation of underreporting across design features.

https://doi.org/10.1093/pan/mpv006 Published online by Cambridge University Press

References

Anderson, Richard G. 2013. Registration and replication: A comment. *Political Analysis* 21:38–39.

Asendorpf, J. B., M. Conner, F. De Fruyt, J. De Houwer, J. J. A. Denissen, K. Fiedler, S. Fiedler, D. C. Funder, R. Kliegl, B. A. Nosek, M. Perugini, B. W. Roberts, M. Schmitt, M. A. G. van Aken, H. Weber, and J. M. Wicherts. 2013. Recommendations for increasing replicability in psychology. *European Journal of Personality* 27:108–19.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *Quarterly Journal of Economics* 127:1755–1812.

Chambers, Christopher D. 2013. Registered reports: A new publishing initiative at *Cortex*. *Cortex* 49:609–10.

Chan, An-Wen, and Douglas G. Altman. 2005. Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *British Medical Journal* 330:753.

Chan, An-Wen, Asbjørn Hróbjartsson, Mette T. Haahr, Peter C. Gøtzsche, and Douglas G. Altman. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Journal of the American Medical Association* 291:2457–65.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science* 345:1502–1505.

———. 2015. Replication data for: Underreporting in political science survey experiments: Comparing questionnaires to published results. http://dx.doi.org/10.7910/DVN/28766 Dataverse [Distributor] V1 [Version], January 21, 2015.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima. 2012. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 5:189–211.

Gelman, Andrew, and Francis Tuerlinckx. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15:373–90.

Gerber, Alan, and Neil Malhotra. 2008. Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science* 3:313–26.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis* 21:1–20.

King, Gary, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, and Héctor Hernández Llamas. 2007. A "politically robust" experimental design for public policy evaluation, with application to the Mexican Universal Health Insurance Program. *Journal of Policy Analysis and Management* 26:479–506.

Laitin, David D. 2013. Fisheries management. *Political Analysis* 21:42–47.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. Van der Laan. 2014. Promoting transparency in social science research. *Science* 343:30–31.

Monogan III, James E. 2013. A case for registering studies of political outcomes: An application in the 2010 House elections. *Political Analysis* 21:21–37.

Moreno, Santiago G., Alex J. Sutton, Erick H. Turner, Keith R. Abrams, Nicola J. Cooper, Tom M. Palmer, and A. E. Ades. 2009. Novel methods to deal with publication biases: Secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *British Medical Journal* 339:b2981.

Rising, Kristin, Peter Bacchetti, and Lisa Bero. 2008. Reporting bias in drug trials submitted to the Food and Drug Administration: Review of publication and presentation. *PLoS Medicine* 5:217.

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. *Psychological Science* 22:1359–1366.

https://doi.org/10.1093/pan/mpv006 Published online by Cambridge University Press