Sciencexpress Reports

Publication bias in the social sciences: Unlocking the file drawer

Annie Franco,[1] Neil Malhotra,[2]* Gabor Simonovits[1]

[1]Department of Political Science, Stanford University, Stanford, CA, USA. [2]Graduate School of Business, Stanford University, Stanford, CA, USA.

*Corresponding author. E-mail: neilm@stanford.edu

We study publication bias in the social sciences by analyzing a known population of conducted studies—221 in total—where there is a full accounting of what is published and unpublished. We leverage TESS, an NSF-sponsored program where researchers propose survey-based experiments to be run on representative samples of American adults. Because TESS proposals undergo rigorous peer review, the studies in the sample all exceed a substantial quality threshold. Strong results are 40 percentage points more likely to be published than null results, and 60 percentage points more likely to be written up. We provide not only direct evidence of publication bias, but also identify the stage of research production at which publication bias occurs—authors do not write up and submit null findings.

Publication bias occurs when "publication of study results is based on the direction or significance of the findings" (*1*). One pernicious form of publication bias is the greater likelihood of statistically significant results being published than statistically insignificant results, holding fixed research quality. Selective reporting of scientific findings is often referred to as the "file drawer" problem (*2*). Such a selection process increases the likelihood that published results reflect Type I errors rather than true population parameters, biasing effect sizes upwards. Further, it constrains efforts to assess the state of knowledge in a field or on a particular topic, since null results are largely unobservable to the scholarly community.

Publication bias has been documented in various disciplines within the biomedical (*3–9*) and social sciences (*10–17*). One common method of detecting publication bias is replicating a meta-analysis with and without unpublished literature (*18*). This approach is limited because much of what is unpublished is unobserved. Other methods solely examine the published literature and rely on assumptions about the distribution of unpublished research by, for example, comparing the precision and magnitude of effect sizes among a group of studies. In the presence of publication bias smaller studies report larger effects in order to exceed arbitrary significance thresholds (*19*, *20*). However, these visualization-based approaches are sensitive to using different measures of precision (*21*, *22*) and also assume outcome variables and effect sizes are comparable across studies (*23*). Finally, methods that compare published studies to "grey" literatures (e.g., dissertations, working papers, conference papers, human subjects registries) may confound strength of results with research quality (*7*). These techniques are also unable to determine whether publication bias occurs at the editorial stage or during the writing stage. Editors and reviewers may prefer statistically significant results and reject sound studies that fail to reject the null hypothesis. Anticipating this, authors may not write up and submit papers that have null findings. Or, authors may have their own preferences to not pursue the publication of null results.

A different approach involves examining the publication outcomes of a cohort of studies, either prospectively or retrospectively (*24*, *25*). Analyses of clinical registries and abstracts submitted to medical conferences consistently find little to no editorial bias against studies with null findings (*26–31*). Instead, failure to publish appears to be most strongly related to authors' perceptions that negative or null results are uninteresting and not worthy of further analysis or publication (*32–35*). One analysis of all IRB-approved studies at a single university over two years found that a majority of conducted research was never submitted for publication or peer-review (*36*).

Surprisingly, similar cohort analyses are much rarer in the social sciences. There are two main reasons for this lacuna. First, there is no process in the social sciences of pre-registering studies comparable to the clinical trials registry in the biomedical sciences. Second, even if some unpublished studies could be identified, there are likely to be substantial quality differences between published and unpublished studies that make them difficult to compare. As noted, previous research attempted to identify unpublished results by examining conference papers and dissertations (*37*) and human subjects registries of single institutions (*36*). However, such techniques may produce unrepresentative samples of unpublished research, and the strength of the results may be confounded with research quality. Conference papers, for example, do not undergo a similar process of peer review as journal articles in the social sciences and therefore cannot be used as a comparison set. This paper is unique in the study of publication bias in the social sciences in that it analyzes a known population of conducted studies and all studies in the population exceed a substantial quality threshold.

We leverage TESS (Time-sharing Experiments in the Social Sciences), an NSF-sponsored program established in 2002 where researchers propose survey-based experiments to be run on nationally representative samples. These experiments typically embed some randomized manipulation (e.g., visual stimulus, question wording difference) within a survey questionnaire. Researchers apply to TESS, which then peer reviews the proposals and distributes grants on a competitive basis (*38*). Our basic approach is to compare the statistical results of TESS experiments that eventually got published to the results of those that remain unpublished.

This analytic strategy has many advantages. First, we have a known population of conducted studies, and therefore have a full accounting of what is published and unpublished. Second, TESS proposals undergo rigorous peer review, meaning that even unpublished studies exceed a substantial quality threshold before they are conducted. Third, nearly all of the survey experiments were conducted by the same, high-quality survey research firm (Knowledge Networks, now known as GfK Custom Research), which assembles probability samples of Internet panelists by recruiting participants via random digit dialing and address-based sampling. Thus, there is remarkable similarity across studies with respect to how they were administered, allowing for comparability. Fourth, TESS requires that studies have requisite statistical power, meaning that the failure to obtain statistically significant results is not simply due to insufficient sample size.

One potential concern is that TESS studies may be unrepresentative of social science research, especially scholarship based on non-experimental data. While TESS studies are clearly not a random sample of the research conducted in the social sciences, it is unlikely that publication bias is *less* severe than what is reported here. The baseline probability of publishing experimental findings based on representative

samples is likely higher than that of observational studies using "off-the-shelf" datasets or experiments conducted on convenience samples where there is lower "sunk cost" involved in obtaining the data. Because the TESS data were collected at considerable expense—in terms of time to obtain the grant—authors should, if anything, be more motivated to attempt to publish null results.

The initial sample consisted of the entire online archive of TESS studies as of January 1, 2014 (*39*). We analyzed studies conducted between 2002 and 2012. We did not track studies conducted in 2013 because there had not been enough time for the authors to analyze the data and proceed through the publication process. The 249 studies represent a wide range of social science disciplines (see Table 1). Our analysis was restricted to 221 studies—89% of the initial sample. We excluded seven studies published in book chapters, and 21 studies for which we were unable to determine the publication status and/or the strength of experimental findings (*40*). The full sample of studies is presented in Table 2; the bolded entries represent the analyzed subsample of studies.

The outcome of interest is the publication status of each TESS experiment. We took numerous approaches to determine whether the results from each TESS experiment appeared in a peer-reviewed journal, book, or book chapter. We first conducted a thorough online search for published and unpublished manuscripts, and read every manuscript to verify that it relied on data collected through TESS and that it reported experimental results (*40*). We then emailed the authors of over 100 studies for which we were unable to find any trace of the study and asked what happened to their studies. We also asked authors who did not provide a publication or working paper to summarize the results of their experiments.

The outcome variable distinguishes between two types of unpublished experiments: those prepared for submission to a conference or journal, and those never written up in the first place. It is also possible that papers with null results may be excluded from the very top journals but still find their way into the published literature. Thus, we disaggregated published experiments based on their placement in top-tier or non-top-tier journals (*40*) (see Table S1 for a list of journal classifications). The results from the majority of TESS studies in our analysis sample have been written up (80%), while less than half (48%) have been published in academic journals.

We also ascertained whether the results of each experiment are described as statistically significant by their authors. We did not analyze the data ourselves to determine if the findings were statistically significant for two main reasons. First, it is often very difficult to discern the exact analyses the researchers intended. The proposals that authors submit to TESS are not a matter of public record, and many experiments have complex experimental designs with numerous treatment conditions, outcome variables, and moderators. Second, what is most important is whether the authors themselves consider their results to be significant, as this influences how they present their results to editors and reviewers, as well as whether they decide to write a paper. Studies were classified into three categories of results: strong (all/most of hypotheses were supported by the statistical tests), null (all/most hypotheses were not supported), and mixed (remainder of studies) (*40*). Approximately 41% of the studies in our analysis sample reported strong evidence in favor of the stated hypotheses, 37% reported mixed results, and 22% reported null results.

There is a strong relationship between the results of a study and whether it was published, a pattern indicative of publication bias. The main findings are presented in Table 3, which is a cross-tabulation of publication status against strength of results. A Pearson chi-squared test of independence is easily rejected [$\chi^2(6) = 80.3$, $P < 0.001$], implying that there are clear differences in the statistical results between published and unpublished studies. While around half of the total studies in our sample were published, only 20% of those with null results appeared in print. In contrast, roughly 60% of studies with strong results and 50% of those with mixed results were published. Although more than 20% of the studies in our sample had null findings, less than 10% of published articles based on TESS experiments report such results. While the direction of these results may not be surprising, the observed magnitude (an approximately 40 percentage point increase in the probability of publication from moving from null to strong results) is remarkably large.

However, what is perhaps most striking in Table 1 is not that so few null results are published, but that so many of them are never even written up (65%). The failure to write up null results is problematic for two reasons. First, researchers might be wasting effort and resources in conducting studies that have already been executed where the treatments were not efficacious. Second, and more troubling, if future researchers conduct similar studies and obtain significant results by chance, then the published literature on the topic will erroneously suggest stronger effects. Hence, even if null results are characterized by treatments that "did not work" and strong results are characterized by efficacious treatments, authors' failures to write up null findings still adversely affects the universe of knowledge. Interestingly, once we condition on studies that were written up, there is no significant relationship between strength of results and publication status (see Table S2).

A series of additional analyses demonstrate the robustness of our results. Estimates from multinomial probit regression models show that studies with null findings are significantly less likely to be written up even after controlling for researcher quality (using the highest quality researcher's cumulative h-index and the number of publications at the time the study was ran), discipline of the lead author, and the date the study was conducted (see online supplementary text and Table S3). Further, the relationship between strength of results and publication status does not vary across levels of these covariates (see online supplementary text and Tables S4 and S5). Another potential concern is that our coding of the statistical strength of results is based on author self-reports, introducing the possibility of measurement error and misclassification. A sensitivity analysis shows that our findings are robust to even dramatic and unrealistic rates of misclassification (see online supplementary text and Figure S1).

Why do some researchers choose not to write up null results? To provide some initial explanations, we classified 26 detailed email responses we received from researchers whose studies yielded null results and did not write a paper (see Table S6). Fifteen of these authors reported that they abandoned the project because they believed that null results have no publication potential even if they found the results interesting personally (e.g., "I think this is an interesting null finding, but given the discipline's strong preference for $p < .05$, I haven't moved forward with it"). Nine of these authors reacted to null findings by reducing the priority of writing up the TESS study and focusing on other projects (e.g., "There was no paper unfortunately. There still may be in future. The findings were pretty inconclusive."). Perhaps most interestingly, two authors whose studies "didn't work out" eventually published papers supporting their initial hypotheses using findings obtained from smaller convenience samples.

How can the social science community combat publication bias of this sort? Based on communications with the authors of many experiments that resulted in null findings, we found that some researchers anticipate the rejection of such papers but also that many of them simply lose interest in "unsuccessful" projects. These findings show that a vital part of developing institutional solutions to improve scientific transparency would be to understand better the motivations of researchers who choose to pursue projects as a function of results.

Few null findings ever make it to the review process. Hence, proposed solutions such as two-stage review (the first stage for the design and the second for the results), pre-analysis plans (*41*), and requirements to pre-register studies (*16*) should be complemented by incentives to not bury insignificant results in file drawers. Creating high-status publication

outlets for these studies could provide such incentives. The movement toward open-access journals may provide space for such articles. Further, the pre-analysis plans and registries themselves will increase researcher access to null results. Alternatively, funding agencies could impose costs on investigators who do not write up the results of funded studies. Finally, resources should be deployed for replications of published studies if they are unrepresentative of conducted studies and more likely to report large effects.

References and Notes

1. K. Dickersin, The existence of publication bias and risk factors for its occurrence. *JAMA* **263**, 1385–1389 (1990). Medline doi:10.1001/jama.1990.03440100097014
2. R. Rosenthal, The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**, 638–641 (1979). doi:10.1037/0033-2909.86.3.638
3. C. B. Begg, J. A. Berlin, Publication bias: A problem in interpreting medical data. *J. R. Stat. Soc. Ser.* **151**, 419–463 (1988). doi:10.2307/2982993
4. J. A. Berlin, C. B. Begg, T. A. Louis, An assessment of publication bias using a sample of published clinical trials. *J. Am. Stat. Assoc.* **84**, 381–392 (1989). doi:10.1080/01621459.1989.10478782
5. P. J. Easterbrook, J. A. Berlin, R. Gopalan, D. R. Matthews, Publication bias in clinical research. *Lancet* **337**, 867–872 (1991). Medline doi:10.1016/0140-6736(91)90201-Y
6. L. McAuley, B. Pham, P. Tugwell, D. Moher, Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? *Lancet* **356**, 1228–1231 (2000). Medline doi:10.1016/S0140-6736(00)02786-0
7. M. Egger, P. Juni, C. Bartlett, F. Holenstein, J. Sterne, How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol. Assess.* **7**, 1–76 (2003). Medline
8. H. R. Rothstein, A. J. Sutton, M. Borenstein, Eds., *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (Wiley, Chichester, U.K., 2006).
9. F. Song, S. Parekh, L. Hooper, Y. K. Loke, J. Ryder, A. J. Sutton, C. Hing, C. S. Kwok, C. Pang, I. Harvey, Dissemination and publication of research findings: An updated review of related biases. *Health Technol. Assess.* **14**, iii, ix–xi, 1–193 (2010). Medline
10. T. D. Sterling, Publications decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* **54**, 30–34 (1959).
11. A. Coursol, E. E. Wagner, Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Prof. Psychol. Res. Pr.* **17**, 136–137 (1986). doi:10.1037/0735-7028.17.2.136
12. A. G. Greenwald, Consequences of prejudice against the null hypothesis. *Psychol. Bull.* **82**, 1–20 (1975). doi:10.1037/h0076157
13. D. Card, A. B. Krueger, Time-series minimum wage studies: A meta-analysis. *Am. Econ. Rev.* **85**, 238–243 (1995).
14. O. Ashenfelter, C. Harmon, H. Oosterbeek, A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour Econ.* **6**, 453–470 (1999). doi:10.1016/S0927-5371(99)00041-X
15. C. Doucouliagos, Publication bias in the economic freedom and economic growth literature. *J. Econ. Surv.* **19**, 367–387 (2005). doi:10.1111/j.0950-0804.2005.00252.x
16. A. Gerber, N. Malhotra, Do statistical reporting standards affect what is published? publication bias in two leading political science journals. *Quart. J. Pol. Sci.* **3**, 313–326 (2008).
17. A. Gerber, N. Malhotra, Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociol. Methods Res.* **37**, 3–30 (2008). doi:10.1177/0049124108318973
18. H. Cooper, L. V. Hedges, J. C. Valentine, Eds., *The Handbook of Research Synthesis and Meta-Analysis* (Russell Sage Foundation, New York, ed. 2, 2009)
19. R. J. Light, D. B. Pillemar, *Summing Up: The Science of Reviewing Research* (Harvard University Press, Cambridge, MA, 1984).
20. M. Egger, G. D. Smith, M. Schneider, C. Minder, Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**, 629–634 (1997). Medline doi:10.1136/bmj.315.7109.629
21. J. P. Ioannidis, T. A. Trikalinos, The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *CMAJ* **176**, 1091–1096 (2007). Medline doi:10.1503/cmaj.060410
22. J. Lau, J. P. Ioannidis, N. Terrin, C. H. Schmid, I. Olkin, The case of the misleading funnel plot. *BMJ* **333**, 597–600 (2006). Medline doi:10.1136/bmj.333.7568.597
23. D. T. Felson, Bias in meta-analytic research. *J. Clin. Epidemiol.* **45**, 885–892 (1992). Medline doi:10.1016/0895-4356(92)90072-U
24. K. Dickersin, in *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, H. R. Rothstein, A. J. Sutton, M. Borenstein, Eds. (Wiley, Chichester, U.K., 2006), ch. 2.
25. K. Dwan, D. G. Altman, J. A. Arnaiz, J. Bloom, A. W. Chan, E. Cronin, E. Decullier, P. J. Easterbrook, E. Von Elm, C. Gamble, D. Ghersi, J. P. Ioannidis, J. Simes, P. R. Williamson, Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLOS ONE* **3**, e3081 (2008). 10.1371/journal.pone.0003081 Medline doi:10.1371/journal.pone.0003081
26. M. L. Callaham, R. L. Wears, E. J. Weber, C. Barton, G. Young, Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *JAMA* **280**, 254–257 (1998). Medline doi:10.1001/jama.280.3.254
27. M. Callaham, R. L. Wears, E. Weber, Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA* **287**, 2847–2850 (2002). Medline doi:10.1001/jama.287.21.2847
28. C. M. Olson, D. Rennie, D. Cook, K. Dickersin, A. Flanagin, J. W. Hogan, Q. Zhu, J. Reiling, B. Pace, Publication bias in editorial decision making. *JAMA* **287**, 2825–2828 (2002). Medline doi:10.1001/jama.287.21.2825
29. A. Timmer, R. J. Hilsden, J. Cole, D. Hailey, L. R. Sutherland, Publication bias in gastroenterological research - a retrospective cohort study based on abstracts submitted to a scientific meeting. *BMC Med. Res. Methodol.* **2**, 7 (2002). Medline doi:10.1186/1471-2288-2-7
30. K. P. Lee, E. A. Boyd, J. M. Holroyd-Leduc, P. Bacchetti, L. A. Bero, Predictors of publication: Characteristics of submitted manuscripts associated with acceptance at major biomedical journals. *Med. J. Aust.* **184**, 621–626 (2006). Medline
31. K. Okike, M. S. Kocher, C. T. Mehlman, J. D. Heckman, M. Bhandari, Publication bias in orthopaedic research: An analysis of scientific factors associated with publication in the Journal of Bone and Joint Surgery (American Volume). *J. Bone Joint Surg.* **90**, 595–601 (2008). Medline doi:10.2106/JBJS.G.00279
32. K. Dickersin, S. Chan, T. C. Chalmers, H. S. Sacks, H. Smith Jr., Publication bias and clinical trials. *Control. Clin. Trials* **8**, 343–353 (1987). Medline doi:10.1016/0197-2456(87)90155-3
33. K. Dickersin, Y. I. Min, C. L. Meinert, Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA* **267**, 374–378 (1992). Medline doi:10.1001/jama.1992.03480030052036
34. K. Dickersin, Y. I. Min, *Online J. Curr. Clin. Trials* **1993**, 50 (1993).
35. R. M. D. Smyth, J. J. Kirkham, A. Jacoby, D. G. Altman, C. Gamble, P. R. Williamson, Frequency and reasons for outcome reporting bias in clinical trials: Interviews with trialists. *BMJ* **342** (jan06 1), c7153 (2011). 10.1136/bmj.c7153 Medline doi:10.1136/bmj.c7153
36. H. Cooper, K. DeNeve, K. Charlton, Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychol. Methods* **2**, 447–452 (1997). doi:10.1037/1082-989X.2.4.447
37. G. V. B. Glass, B. McGaw, M. L. Smith, *Meta Analysis in Social Research* (Sage, Beverly Hills, CA, 1981).
38. The rate at which research-initiated proposals are approved by the peer reviewers engaged by TESS is provided in the supplementary materials.
39. TESS archive; www.tessexperiments.org.
40. Materials and methods are available as supplementary material on *Science* online.
41. K. Casey, R. Glennerster, E. Miguel, Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *Q. J. Econ.* **127**, 1755–1812 (2012). doi:10.1093/qje/qje027
42. A. W. Harzing, Publish or Perish, available from http://www.harzing.com/pop.htm (2007).

Acknowledgments: Data and replication code are available on GitHub (doi:

10.5281/zenodo.11275). All authors contributed equally to all aspects of the research. No funding was required for this article. The authors declare no conflicts of interest. We thank seminar participants at the 2014 Annual Meeting of the Midwest Political Science Association, the 2014 Annual Meeting of the Society for Political Methodology, the 2014 West Coast Experiments Conference, Stanford University, and U.C. San Diego. We thank Christopher McConnell and Stacy Liu for valuable research assistance.

Supplementary Materials
www.sciencemag.org/content/science.1255484/DC1
Materials and Methods
Supplementary Text
Fig. S1
Tables S1 to S7
Reference (*42*)

1 May 2014; accepted 14 August 2014
Published online 28 August 2014
10.1126/science.1255484

Table 1: Distribution of studies across years and disciplines. Note: Field coded based on the affiliation of the first author. "Other" category includes: Business, Computer Science, Criminology, Education, Environmental Studies, Journalism, Law, and Survey Methodology.

Year	Communication	Economics	Political Science	Public Health	Psychology	Sociology	Other	Total
2002	0	0	1	0	0	0	0	1
2003	0	1	4	0	6	2	1	14
2004	0	2	9	1	5	0	0	17
2005	2	2	13	0	10	7	1	35
2006	3	1	12	1	9	6	0	32
2007	0	0	5	0	3	2	0	10
2008	2	0	11	1	4	2	1	21
2009	0	0	12	1	8	2	3	26
2010	3	3	22	0	5	6	2	41
2011	2	0	19	1	9	6	2	39
2012	1	1	5	1	1	3	1	13
Total	13	10	113	6	60	36	11	249

Table 2. Cross-tabulation between statistical results of TESS studies and their publication status. *Note:* Entries are counts of studies by publication status and results. Bolded entries indicate observations included in the final sample for analysis (*40*). Results are robust to the inclusion of book chapters (see Table S7).

	Unpublished, Not written	Unpublished, Written	Published	Book chapter	Missing	Total
Null results	31	7	10	1	0	49
Mixed results	10	32	40	3	1	86
Strong results	4	31	56	1	1	93
Missing	6	1	0	2	12	21
Total	51	71	106	7	14	249

Table 3. Cross-tabulation between statistical results of TESS studies and their publication status (column percentages reported). Pearson χ^2 test of independence: $\chi^2(6) = 80.3$, $P < 0.001$.

	Null	Mixed	Strong
Not written	64.6%	12.2%	4.4%
Written but not published	14.6	39.0	34.1
Published (non-top-tier)	10.4	37.8	38.5
Published (top-tier)	10.4	11.0	23.1
Total	100.0	100.0	100.0