Supplementary Materials for

Title: Publication Bias in the Social Sciences: Unlocking the File Drawer

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

Affiliations:

[1]Department of Political Science, Stanford University, Stanford, CA

[2]Graduate School of Business, Stanford University, Stanford, CA

*Correspondence to: neilm@stanford.edu

This PDF file includes:

Materials and Methods
Supplementary Text
Tables S1 to S7
Figure S1

Other Supplementary Materials for this manuscript includes the following:

N/A

Materials and Methods

Background on TESS

TESS (Time-sharing Experiments in the Social Sciences) is a National Science Foundation sponsored program where researchers propose survey-based experiments to be run on representative samples of the U.S. adult population at no cost to the researchers. Proposals undergo peer review to determine whether the study is run. Accepted studies are administered over the Internet to a panel of survey respondents assembled by GfK Custom Research, a market research firm. A requirement is that all proposed studies include randomized experiments embedded within the survey. For example, respondents can be randomly assigned to receive different types of stimuli (e.g., news articles to read, videos to watch, alternative question wordings) before answering a series of survey questions. TESS also funds designs that leverage within-subjects experiments. More information about the program can be found at www.tessexperiments.org.

Initial Sample

Our initial sample consisted of the entire online archive of TESS studies as of January 1, 2014, or the 249 studies conducted by TESS between 2002 and 2012 available on the site at that time. The materials from an additional number of studies conducted in 2012 were released online after our data collection efforts ended and are excluded from our analyses.

Publication Status

Our outcome of interest is a study's publication status. To determine whether a study was published, we first performed various searches on Google Scholar and ISI Web of Science for: (1) the name of the study (as well as key words from the study title); (2) the authors' names; (3) the words "TESS" or "Time-sharing Experiments in the Social Sciences." We also examined the vitaes of scholars who received TESS grants and reviewed their published papers to see if the TESS experiments had appeared in print.

After identifying articles that potentially included the results of each study, we read each one to verify that the results relied on data collected through TESS and that they report experimental findings (i.e., differences in outcome variables between conditions). Next, we attempted to collect unpublished manuscripts based on the TESS studies. While we were able to track down many working papers based on the studies, we were still left with more than a hundred studies for which we could not find any trace of their publication status online.

To learn more information about these projects and to make sure we had not missed some published articles in our searches, we emailed all the authors of these seemingly unpublished studies and asked about the publication status of their papers. Our email communications also asked whether, if unpublished, the study had been written up as a working or conference paper and submitted to a journal, or if no paper was ever written. This allows us to distinguish two types of unpublished studies: (1) those prepared for submission to conferences or journals; and (2) those never even written up in the first place.

Additionally, we coded whether the studies were published in a top-tier journal or a non-top-tier journal. We identified top-tier journals as those ranked in the top four in their respective categories according Google Scholar's 2014 h5-index (see Table S1).

Strength of Results

For each study we were able to locate, two of the authors of the present paper coded the results into three categories: (1) strong findings; (2) mixed findings; and (3) null results. For published studies and working papers we based the coding on how the PIs themselves pitched and framed the results, relying mostly on the description of the experimental findings as reported in the abstract, results section, tables and figures, and conclusion. For studies where we could not access write-ups (either because they were never written up or because we could not obtain the papers from the authors) but we managed to contact the authors, we relied on the authors' own summaries of the results. In those cases where we could not contact the authors, we relied on the summaries on the TESS website. TESS asks authors to report summaries of their findings one year after the data are collected, but not all authors complete this report.

While ascertaining the strength of results is obviously a somewhat subjective exercise, we sought to define coding rules that made classification the most exact possible. In particular, we coded results as "strong" if all or most of the expectations (hypotheses) appearing in the article were supported by the statistical tests. We categorized results as "null findings" if most expectations were not supported. The results of the remaining studies were coded as "mixed." The "mixed" studies were often characterized by studies where some hypotheses were supported while others were not, where significant treatment effects were obtained for some outcome variables and not others, and where there was no significant treatment effect in a main sample but significant results were detected in subgroups. For studies that did not specify any hypotheses we assumed that the authors expected some mean differences across dependent variables in the different experimental conditions. In the case of studies that hypothesized no differences across experimental groups, we still coded results as "null findings" to guard against the possibility of post hoc hypothesis construction (there were few studies of this type, and our results are not sensitive to this coding decision).

We followed the same classification rule with studies that we learned about through direct communication with the author(s) or through the summaries appearing on the TESS website. Email communications about the results of studies were mostly easy to classify as well. Inter-coder reliability was high (89%). In cases of disagreement, all three authors of the present paper discussed the study and agreed upon a coding.

Additional Variables

We collected additional data on the studies and their authors to use as control variables in our multivariate models. First, using data available from the TESS website, we recorded the exact date when the survey experiment was administered. Using this date, we defined the "age" of each study as the number of days elapsed between the first day of the fielding period and the date we carried out the statistical analysis (1, August, 2014). For two studies only the year the survey was administered was reported; we coded the date of administration as 15, June for these two studies.

We also collected several measures that proxy for researcher quality. Using the Publish or Perish tool (*42*) we recorded the current h-indices of the PIs of each study, as well as their total number of publications at the time the TESS study was conducted. The publications identified by Publish or Perish were verified against each author's CV or Google Scholar profile. Our statistical analyses include the maximum value of each measure among the PIs of a specific study.

Finally, we collected data on the disciplinary affiliation of each investigator. We identified disciplines by considering each researcher's current department, his/her recent publications, the fields in which he/she earned their terminal degrees, and his/her stated research interests. We then coded each study as belonging to the field of the first author of the study. For 78% of studies with multiple authors, the authors belonged to the same discipline. Thus, our decision to use the affiliation of the first author was immaterial for most observations.

Excluded and Missing Data

Our analysis consists of 221 studies—89% of the full sample of 249—after excluding the studies for which we were unable to determine publication status and/or strength of results, as well as studies for which these categories were not applicable.

First, two studies were excluded because we learned that the authors did not list any hypotheses that compared outcomes across experimental conditions, which is the definition of "experimental results" used in the present paper.

Second, six studies were excluded because our communications with the authors revealed that they either have not analyzed the data (five studies) or did not recall what the results were (one study).

Third, 14 studies were excluded because we were unable to find a manuscript online or contact the authors. Out of these 14 studies, we were able to code the strength of the results for two studies based on the descriptions provided on the TESS website, but could not determine publication status through author communication or online searches. For the remaining 12 studies, we were unable to collect information about either publication status or the results.

Finally, seven studies published as book chapters in edited volumes were also excluded from our analysis since peer review standards for these types of publications are different than those of academic journals. Our results are unaffected by whether we categorize book chapters as published (top-tier or non-top-tier), unpublished, or treat them as a separate category (Table S7 replicates Table 3 in the main text with these alternative coding schemes).

Supplementary Text

Robustness Check: Models with Controls

To assess the robustness of the association between the strength of results and publication status we estimated multinomial probit (MNP) models including potential confounders as covariates. This statistical method models the conditional probabilities of the observed outcomes (here, published, unpublished but written, and unwritten) as a function of covariates. Among possible modeling approaches MNP was preferred because it does not require the assumption that unobserved variables affecting the probabilities of the modeled outcomes are independent (this would be a very implausible assumption in our case).

The results reported in Table S4 confirm the findings in Table 3 based on the simple approach of cross-tabulating results and publication status. The stability of our coefficients of interest across specifications suggests that the relationship between results and publication status is not due to some plausible alternative explanation (such as, better researchers being more likely to find significant results, and also being better able to publish).

The results reported in Table S4 show that the association between publication status and the strength of results is not conditional on the date the study was conducted; nor does the relationship differ across "high quality" and "low quality" scholars (as measured by the maximal author's h-index) or "more productive" or "less productive" scholars (as measured by the maximal author's number of published articles at the time the study was conducted). Overall these estimates (along with the ones reported in Tables S5) suggest that, holding other factors constant, our results do not vary according to discipline, author quality, or the date the study was administered.

Robustness Check: Sensitivity to Misclassification

We conducted a sensitivity analysis to bound the results from Table 3 in the presence of misclassification. Because the initial results were driven by the cell representing null studies that were never written up, we collapsed the publication status variable into a 2x2 contingency table reflecting the main variables of interest (null vs. non-null results; written vs. unwritten studies). We considered two types of misclassification: (1) unwritten studies coded as having null findings but which actually had non-null findings; (2) written studies coded as having non-null findings but which actually had null findings. We then calculated how many studies would need to be misclassified in order to overturn our results. We find that there would have to be a dramatic amount of miscoding for our results to be overturned, suggesting that our findings are robust to measurement error in the strength of results variable. As shown in Figure S1, over two-thirds of unwritten studies coded as having null findings would have to be miscoded to overturn our results. Over 45% of written studies coded as having non-null findings would have to be miscoded to overturn our results. In terms of joint misclassifications, about one-third of unwritten studies with null findings *and* over 20% of written studies with null findings would have to be miscoded to push our results beyond statistical significance.

<insert page break then Fig S1 here>

Fig. S1.

Sensitivity of Pearson chi-squared test of independence in Table 3 to misclassification of TESS studies.

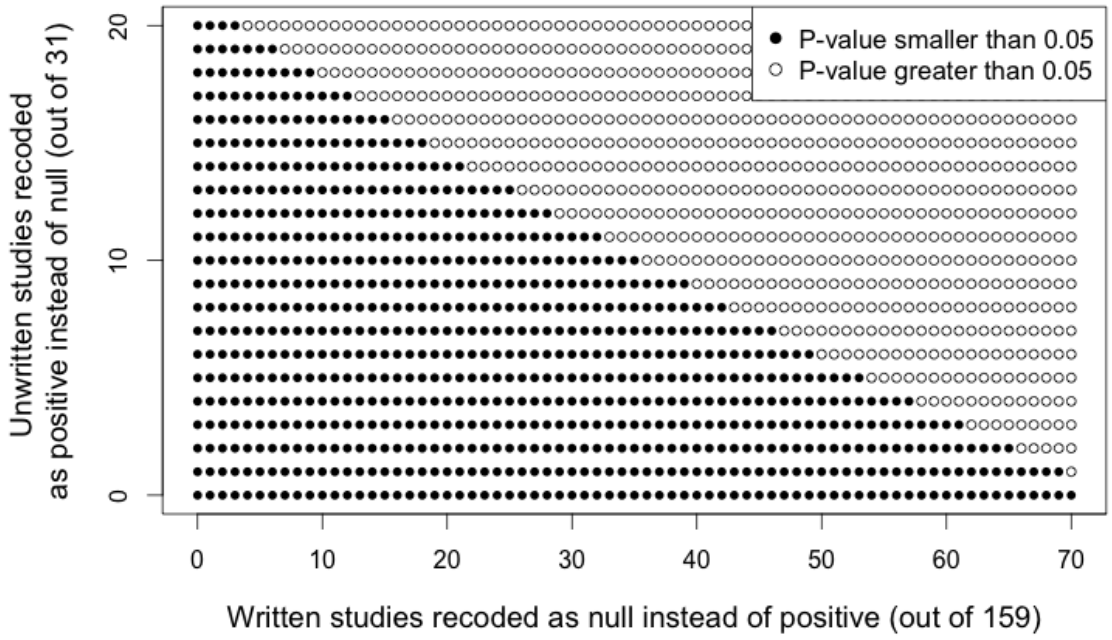

Table S1.

Distribution of published TESS experiments across disciplines.

COMMUNICATION (16)	**POLITICAL SCIENCE (34)**
J. of Communication (1)	*American J. of Political Science (4)*
Public Opinion Quarterly (9)	*American Political Science Review (1)*
Communication Quarterly (1)	*J. of Politics (6)*
Communication Studies (1)	American Politics Research (1)
International J. of Internet Science (1)	British J. of Political Science (1)
International J. of Public Opinion Research (1)	Comparative Politics (1)
Journalism and Mass Comm. Quarterly (1)	Election Law J. (1)
Western J. of Communication (1)	J. of Experimental Political Science (1)
GENERAL (6)	Political Analysis (1)
PNAS (2)	Political Behavior (5)
PLoS ONE (1)	Political Communication (1)
Social Science Quarterly (3)	Political Psychology (3)
LAW (5)	Political Research Quarterly (2)
J. of Empirical Legal Studies (3)	Political Science Quarterly (1)
Minnesota Law Review (1)	Presidential Studies Quarterly (1)
New York University Law Review (1)	PS: Political Science and Politics (1)
OTHER (9)	Public Choice (1)
Health Affairs (3)	Quarterly J. of Political Science (1)
American J. of Education (1)	State Politics & Policy Quarterly (1)
Climactic Change (1)	**ECONOMICS/BUSINESS (5)**
Evolution and Human Behavior (1)	*American Economic Review (2)*
J. for the Scientific Study of Religion (1)	American Economic Review P&P (1)
Secularism and Nonreligion (1)	J. of Consumer Affairs (1)
Social Justice Research (1)	J. of Public Economics (1)
SOCIOLOGY (8)	**PSYCHOLOGY/SOCIAL PSYCHOLOGY (16)**
Journal of Marriage and Family (1)	*J. of Experimental Social Psychology (2)*
Analyses of Social Issues and Public Policy (1)	*J. of Personality and Social Psychology (2)*
J. of Family Issues (1)	*Personality and Individual Differences (1)*
Social Networks (1)	*Personality and Social Psychology Bulletin (1)*
Social Problems (1)	*Psychological Science (1)*
Sociological Methods and Research (1)	American J. of Psychiatry (1)
Sociological Spectrum (1)	Australian J. of Psychology (1)
Survey Research Methods (1)	Current Research in Social Psychology (1)
INTERNATIONAL RELATIONS (7)	Emotion (1)
International Organization (2)	J. of Family Psychology (1)
Foreign Policy Analysis (2)	Org. Behavior & Human Decision Processes (1)
International Studies Quarterly (1)	Personal Relationships (1)
Security Studies (1)	Rehabilitation Psychology (1)
Terrorism and Political Violence (1)	Social Psychology Quarterly (1)
TOTAL	**71 JOURNALS, 106 ARTICLES**

Note: Top-tier journals in italics. Number of articles per journal and field in parentheses.

Table S2.

Cross-tabulation between statistical results of TESS studies and their publication status conditional on existence of a written report. Column percentages reported.

	Null	Mixed	Strong
Written but not published	41.2%	44.4%	35.6%
Published (non-top-tier)	29.4	43.1	40.2
Published (top-tier)	29.4	12.5	24.1
Total	100.0	100.0	100.0

Pearson chi-squared test of independence: χ^2 (4) = 5.0, p = .29

Table S3.

Multinomial probit regressions predicting publication status as a function of strength of results, age of study, researcher quality, and discipline.

Dependent variable: Publication status	(1)		(2)	
	Published	Unwritten	Published	Unwritten
Mixed results	-0.32	0.51	-0.17	0.95**
	(0.28)	(0.40)	(0.30)	(0.46)
Null results	-0.26	2.53***	-0.58	2.89***
	(0.39)	(0.44)	(0.44)	(0.51)
Age of study (100 days)	——	——	0.09***	0.05**
			(0.02)	(0.02)
h-index	——	——	0.01	0.03
			(0.02)	(0.02)
Number of publications	——	——	-0.00	-0.01
			(0.00)	(0.00)
Political Science	——	——	-0.59	-0.92*
			(0.39)	(0.49)
Psychology	——	——	0.26	1.21**
			(0.50)	(0.55)
Sociology	——	——	-0.60	0.60
			(0.54)	(0.62)
Constant	0.51***	-1.35***	-1.22**	-2.92***
	(0.19)	(0.31)	(0.50)	(0.70)
Probability difference between null results and strong results	-40.7%	60.2%	-37.3%	48.2%
Observations	221	221	221	221

Robust standard errors in parentheses. Statistical tests are two-sided z-tests based on the asymptotic distribution of maximum likelihood estimates. *** $p<0.01$, ** $p<0.05$, * $p<0.10$. The omitted outcome is unpublished but written study. Omitted categories are "Strong results" and "Other disciplines." Predicted probabilities calculated by varying strength of results and holding all other variables at medians or modes.

Table S4.

Multinomial probit regressions predicting publication status and testing for conditional effects of strength of results by age of study and researcher quality.

	(1)		(2)		(3)	
Dependent variable: Publication status	Published	Unwritten	Published	Unwritten	Published	Unwritten
Mixed results	0.17	0.17	-0.07	0.26	-0.06	0.29
	(0.82)	(1.52)	(0.43)	(0.63)	(0.37)	(0.52)
Null results	-1.56	4.61***	-1.01*	2.78***	-0.44	2.52***
	(1.64)	(1.58)	(0.61)	(0.72)	(0.54)	(0.61)
Age (100 days)	0.09***	0.06	0.09***	0.05**	0.09***	0.05**
	(0.03)	(0.05)	(0.02)	(0.02)	(0.02)	(0.02)
h-index	0.01	0.03	0.01	0.02	0.01	0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Number of publications	-0.00	-0.01	-0.00	-0.01	-0.00	-0.02*
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
Political Science	-0.57	-1.08**	-0.59	-0.91*	-0.61	-0.87*
	(0.39)	(0.48)	(0.40)	(0.49)	(0.40)	(0.49)
Psychology	0.30	1.18**	0.30	1.13**	0.25	1.27**
	(0.51)	(0.55)	(0.51)	(0.54)	(0.51)	(0.55)
Sociology	-0.56	0.47	-0.58	0.52	-0.61	0.57
	(0.54)	(0.64)	(0.54)	(0.60)	(0.55)	(0.61)
Weak results x Age (100 days)	-0.02	0.03	———	———	———	———
	(0.04)	(0.06)				
Null results x Age (100 days)	0.02	-0.07	———	———	———	———
	(0.05)	(0.06)				
Mixed results x h-index	———	———	-0.01	0.03	———	———
			(0.02)	(0.02)		
Null results x h-index	———	———	0.01	0.00	———	———
			(0.02)	(0.02)		
Mixed results x Number of publications	———	———	———	———	0.00	0.02**
					(0.00)	(0.01)
Null results x Number of publications	———	———	———	———	0.00	0.01
					(0.00)	(0.01)
Constant	-1.41**	-3.25***	-1.24**	-2.68***	-1.29**	-2.53***
	(0.65)	(1.25)	(0.52)	(0.70)	(0.52)	(0.65)
Observations	221	221	221	221	221	221

Robust standard errors in parentheses. Statistical tests are two-sided z-tests based on the asymptotic distribution of maximum likelihood estimates. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. The omitted outcome is unpublished but written study. Omitted categories are "Strong results" and "Other disciplines."

Table S5.

Cross-tabulation between statistical results of TESS studies and their publication status (by discipline). Column percentages reported.

	Political Science and IR (n=113)			Psychology (n=60)		
	Null	Mixed	Strong	Null	Mixed	Strong
Not written	42.1%	4.6%	0.0%	85.7%	30.8%	12.0%
Written, but not published	31.6	45.5	45.2	0.0	23.1	12.0
Published (non-top-tier)	15.8	36.4	28.6	0.0	38.5	44.0
Published (top-tier)	10.5	13.6	26.2	14.3	7.7	32.0
Total	100.0	100.0	100.0	100.0	100.0	100.0
	Pearson chi-squared test of independence: χ^2 (6) = 31.5, p < 0.001			Pearson chi-squared test of independence: χ^2 (6) = 25.4, p < 0.001		

	Sociology (n=36)			Other (n=40)		
	Null	Mixed	Strong	Null	Mixed	Strong
Not written	83.3%	30.0%	7.7%	66.7%	6.7%	0.0%
Written, but not published	16.7	30.0	38.5	0.0	40.0	36.4
Published (non-top-tier)	0.0	30.0	53.9	22.2	46.7	45.5
Published (top-tier)	0.0	10.0	0.0	11.1	6.7	18.2
Total	100.0	100.0	100.0	100.0	100.0	100.0
	Pearson chi-squared test of independence: χ^2 (6) = 13.6, p = 0.03			Pearson chi-squared test of independence: χ^2 (6) = 18.5, p = 0.005		

Table S6.

Classification of e-mail correspondences of researchers who did not produce a written report for studies with null findings.

Plans for Project	Content of Email
Abandoned project (n=15)	"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", "complete failure, not under review anywhere", "data were buried in the graveyard of statistical findings", "there were no publishable result", "never published the study, it was mostly a disappointing wash", "The unfortunate reality of the publishing world are that null effects do not tell a clear story", "we determined that there was nothing there that we could publish in a professional journal", "results have not been published, as they missed statistical significance", "Unfortunately, these data were never published. They were very confusing", "never published, definitely disappointed to not see any major effects", "The findings weren't significant and weren't written up", "didn't have much luck in getting these results out to publication", "embarrassing, but I never published the data", "We do not have a draft of the results. Although we had some pilot data that supported our results, the TESS study produced null results.", "We did not pursue this any further or publish it (our problem was that the online participants did not believe [the treatment])"
Delayed project (n=9)	"There is no paper as of yet. The hypotheses of the study were not confirmed", "our experimental manipulation was a bust, no paper yet," "still analyzing those data [2011]", "I never got to do it due to a combination of time constraints, some loss of interest on the topic, and lack of earth-shattering results", "paper has been delayed as I have been working on other projects in the meantime", "this got put on the back burner. We're actually just now planning to do the analyses some time in the next few weeks--but nothing yet", "results were disappointing and the associated research was mothballed", "I have yet to do anything with the results", "there was no paper unfortunately. There still may be in future. The findings were pretty inconclusive"
Substituted experiment (n=2)	"the study was unproductive… nothing came of it. The non-TESS version of the same study, in which we used a student sample, did yield fruit. We have a piece in JOP by that title", "I have attached the paper here. Although the TESS results didn't make it in the paper (we had an unanticipated issue with a lot of participants not seeming to understand [the treatment]"

Note: Quotations from e-mail correspondences with researchers.

Table S7.

Cross-tabulation between statistical results of TESS studies and their publication status (including book chapters). Column percentages reported.

	Book chapters coded as a new category			**Book chapters coded as unpublished**		
	Null	Mixed	Strong	Null	Mixed	Strong
Not written	63.3%	11.8%	4.4%	63.3%	11.8%	4.4%
Written, but not published	14.3	37.7	33.7	16.3	41.2	34.8
Published (non-top-tier)	10.2	36.5	38.0	10.2	36.5	38.0
Published (top-tier)	10.2	10.6	22.8	10.2	10.6	22.8
Book chapter	2.0	3.5	1.1	—	—	—
Total	100.0	100.0	100.0	100.0	100.0	100.0
	Pearson chi-squared test of independence: χ^2 (8) = 81.7, p < 0.001			Pearson chi-squared test of independence: χ^2 (6) = 80.6, p < 0.001		
	Book chapters coded as published (non-top-tier)			**Book chapters coded as published (top-tier)**		
	Null	Mixed	Strong	Null	Mixed	Strong
Not written	63.3%	11.8%	4.4%	63.3%	11.8%	4.4%
Written, but not published	14.3	37.7	33.7	14.3	37.7	33.7
Published (non-top-tier)	12.2	40.0	39.1	10.2	36.5	38.0
Published (top-tier)	10.2	10.6	22.8	12.2	14.1	23.9
Total	100.0	100.0	100.0	100.0	100.0	100.0
	Pearson chi-squared test of independence: χ^2 (6) = 80.6, p < 0.001			Pearson chi-squared test of independence: χ^2 (6) = 78.6, p < 0.001		