Political Analysis, 9:4

Testing for Publication Bias in Political Science

Alan S. Gerber, Donald P. Green, and David Nickerson
Department of Political Science, Yale University,
New Haven, CT 06520-8301
e-mail: alan.gerber@yale.edu

If the publication decisions of journals are a function of the statistical significance of research findings, the published literature may suffer from "publication bias." This paper describes a method for detecting publication bias. We point out that to achieve statistical significance, the effect size must be larger in small samples. If publications tend to be biased against statistically insignificant results, we should observe that the effect size diminishes as sample sizes increase. This proposition is tested and confirmed using the experimental literature on voter mobilization.

1 Introduction

THE DEARTH OF insignificant findings in journals reflects the behavior of both researchers and journal editors. Editors and referees look askance at papers that report insignificant findings (Mahoney 1977), and their reputation for doing so creates the "file-drawer problem"—researchers elect not to submit their findings when their research fails to reject the null hypothesis (Iyengar and Greenhouse 1988; Greenwald 1975).

If articles that do not reject the null hypothesis tend to go unpublished, surveys of published research will create a distorted impression about effect size. To achieve statistical significance, studies with a small sample size require larger estimated effects than those with large samples. Publication bias against statistically insignificant results is therefore a directly testable proposition. One can detect the presence of publication bias by plotting the size of the estimated effect by the sample size (Begg 1985, 1994). For one-tailed tests, the smaller the sample size, the larger the published effect size (Light and Pillemer 1984).

Does publication bias inhabit political science? Although the phenomenon has been well documented in other fields such as psychology (e.g., Coursol and Wagner 1986), medical sciences (e.g., Simes 1986; Begg and Berlin 1988; Dickersin 1990), and economics (e.g., DeLong and Lang 1992), the only extended discussion of publication bias in political science is by Lee Sigelman (1999, p. 206) who argues that small sample size is symptomatic of poor methodology. According to this explanation, what may appear to be bias toward statistical significance may instead be an innocuous process whereby methodologically

Authors' note: We are grateful for the useful comments from the three anonymous referees. We are also grateful to the Smith Richardson Foundation and the Institution for Social and Policy Studies at Yale, which helped fund this research, but bear no responsibility for its content.

deficient studies are denied publication. Sigelman's suggestion has an observable implication: published studies based on small samples should be well-executed studies, but there should be no tendency for studies based on small samples to show unusually large effects. This is an empirical question, and in this paper we offer a test of publication bias. After illustrating how in theory publication biases can alter the balance of findings in a research literature, we examine the experimental work on voter turnout, which contains studies with widely varying sample sizes. This literature strongly suggests the presence of publication bias.

2 Detecting Publication Bias

Before delving into the empirical findings on voter mobilization, let us step back and think more generally about similar research situations. Consider the case where the dependent variable is binary (e.g., vote or not) and the natural hypothesis test is a one-tailed test (e.g., the alternative hypothesis states that get-out-the-vote reminders increase voting rates).[1] To simplify the presentation, we describe the analysis of data produced by an experiment measuring the effect of some political contact on the probability that a subject votes. After performing an experiment, the researcher performs a one-sided test of the null hypothesis that there is no treatment effect. The hypothesis is rejected when

$$R = t^* \frac{\sqrt{N}}{\sigma\gamma} > Z_\alpha \tag{1}$$

where Z_α is the value of the α percentile level for the standard normal distribution, $t^* = P_\mathrm{T} - P_\mathrm{C}$ (the treatment effect), P_T is the proportion voting in the treatment group, P_C is the proportion voting in the control group, N is the total sample size, γ equals the product of the inverse of the fraction of the total sample in the treatment group times the inverse of the fraction of the total sample in the control group, and σ is the population standard deviation.[2]

Suppose that a journal review process is captured by a function: the probability that the paper is accepted equals $f(R)$, where f is an increasing function. A special case of $f(R)$ occurs when the publication process can be summarized by a "cutoff rule," whereby publication occurs only if R is greater than some cutoff level (e.g., the 5% significance level). Use of a "cutoff rule" has two important consequences.

First, the expected value of the published treatment effect is always greater than the true treatment effect. This follows from the properties of a truncated normal distribution. Suppose that some variable X is distributed $N(\mu, \sigma^2)$. For any constant A, the expected value of X, such that $X > A$, is

$$E[X \mid X > A] = \mu + \lambda(\alpha)\sigma \tag{2}$$

where $\alpha = (A - \mu)/\sigma$, and, using standard notation for the normal density and cumulative distribution functions, $\lambda(\alpha) = \varphi(\alpha)/(1 - \Phi(\alpha))$, which is strictly greater than zero. Since using a cutoff rule amounts to truncating a normal sampling distribution from below, Eq. (2) shows that estimates will be biased upward.

[1] When articles are published based on the results of one-sided tests, *smaller* effect sizes will go unpublished; when two-sided tests are used, studies will go unpublished when the *absolute value* of their effect sizes is small.

[2] Equation (1) incorporates the approximation that the sampling distribution of t^* is normal, with a variance equal to σ^2/N.

Table 1 The effect of a "5% significance level" rule on reported treatment effects[a]

True effect	*Expected value of reported effect (%)*							
	N = 50	*N = 100*	*N = 200*	*N = 500*	*N = 1000*	*N = 5000*	*N = 10,000*	*N = 30,000*
10%	30.8	22.5	16.7	12.1	10.4	10.0	10.0	10.0
5%	29.9	21.4	15.4	10.2	7.6	5.1	5.0	5.0
1%	29.3	20.7	14.7	9.4	6.7	3.1	2.2	1.4

[a]The examples assume that the sample is divided equally into treatment and control groups and that the population turnout probability for the control group is 0.5. The cell values are the expected value of the reported treatment effect for each sample size and true effect size.

Second, the expected value of the published treatment effect decreases as the sample size increases. If the publication process is a cutoff rule, then results are reported in the literature only if $R > Z_\alpha$, which implies that reported literature is a truncated portion of the entire sampling distribution of t^*. Using Eqs. (1) and (2), the expected value of $t^* \mid R > Z_\alpha$ equals

$$E[t^* \mid R > Z_\alpha] = t + \lambda(x)\frac{\sigma\gamma}{\sqrt{N}} \tag{3}$$

where $x = Z_\alpha - (tN^{1/2}/\sigma\gamma)$, and t is the true treatment effect. Differentiating Eq. (3) with respect to the sample size (N) yields

$$\frac{d(E[t^* \mid R > Z_\alpha])}{d(N)} = \frac{-N^{-3/2}}{2}[\lambda(x)]\left[(\lambda(x) - x)\frac{t\sqrt{N}}{\sigma\gamma} + 1\right] \tag{4}$$

Since $\lambda(\alpha)(\lambda(\alpha) - \alpha) > 0$ for any α,[3] Eq. (4) is strictly negative for any positive N. This shows that as the sample size increases, the expected value of t^* decreases.

Table 1 uses Eq. (3) to construct examples illustrating how a cutoff rule biases published estimates of treatment effects. The tables report the expected value of published results for a variety of treatment effects and sample sizes. Table 1 illustrates how published treatment effects are biased when the cutoff rule is the 5% significance level (one-sided test). We see that, for small samples, the published treatment effects dramatically overestimate the true treatment effect.

The overestimation is especially pronounced when the true effect is very small. Suppose that the true effect is 0.01 (the treatment causes a 1-percentage-point increase in the probability that the subject votes). When the sample is 50, the expected reported treatment effect is just under 0.3, 30 times the true effect. When the sample size is 100, the expected published treatment effect is 0.207, or about 20 times the true effect. Larger sample sizes reduce, but do not eliminate, the large upward bias in reported results. Significant overestimation of the treatment effect is present even as the sample size approaches 500. In that case the expected published treatment effect declines to 0.094, which is still an 800% overestimation of the true effect. For a sample size of 10,000, the cutoff rule of 5% significance will cause the average reported estimate to be more than double the true parameter of 0.01. For very large samples, the bias shrinks and becomes relatively unimportant. The pattern is similar for treatment effects larger than 0.01. When the sample size is small, the published literature reports effects that are several times larger than the true effect. As the sample size increases,

[3]See Greene (1997, pp. 951–952) for a statement of this and other properties of the truncated normal distribution.

Table 2 The probability that effects obtained by a replication study are smaller than those in the published literature[a]

True effect	$N = 50$	$N = 100$	$N = 200$	$N = 500$	$N = 1000$	$N = 5000$	$N = 10{,}000$	$N = 30{,}000$
10%	.93	.89	.83	.68	.55	.50	.50	.50
5%	.96	.95	.93	.88	.79	.53	.50	.50
1%	.98	.98	.97	.97	.96	.93	.88	.76

[a]The cell entries show the probability that a replication study with a sample size equal to N and the true treatment effect listed in the first column produces a result smaller than the average published finding under similar conditions using the cutoff rule examined in Table 1.

the extent of bias declines, but the bias caused by using a cutoff rule for publication is relatively small only when the sample size is in the thousands.

The specter of publication bias calls into question the diagnostic value of small studies. Looking down the columns in Table 1, we observe the startling finding that, in the presence of publication bias, *when the sample size is small, the magnitude of the actual effects has almost no effect on the expected value of the published effects*. Stated somewhat differently, the magnitude of published results may say more about the publication process than about the causal process under investigation. As a practical matter, note that the sample sizes required to mitigate publication bias tend to exceed the sample sizes commonly used in political science, particularly in experimental studies.[4]

The final implication concerns replications of published studies when there is publication bias. For the case analyzed here (where a one-sided hypothesis test is used), the size of the treatment effect uncovered in a replication study will be smaller than the average published effect, and will be much lower when the typical sample size in the existing literature is small. Table 2 provides the probability that a single replication produces a result as large as the average of the published literature under the conditions analyzed in Table 1. Table 2 suggests that, unless the true effect is large, or sample sizes are large, it is very likely that replication studies will find smaller effects than those reported in the literature.

3 Data

To gauge the presence of publication bias empirically, we assembled a group of research publications on the effectiveness of voter mobilization campaigns. This literature was selected because it is one of the few in political science to estimate parameters using randomized experimental design. Thus, the vagaries of model specification that cloud meta-analyses of nonexperimental data analysis are not at issue. Another feature of this literature also recommends it. The kinds of manipulations used in these studies are all fairly similar. While each mobilization campaign contacted voters using somewhat different appeals, together they form a set of relatively similar interventions. Although the studies span several decades, one finds telling variations in sample size among studies conducted at the same point in time (cf., Miller et al. 1981; Adams and Smith 1980).

Table 3 presents the results of published voter mobilization studies. Recall the two general predictions from our model of publication bias: The published literature will tend

[4]The pattern reported in Table 1 generalizes to more relaxed publication rules. For example, when the true treatment effect is 0.01 and the sample size is 100, publishing only those studies which show positive coefficients will lead to an average reported estimate of 0.084, a 740% overestimation of the true effect.

Table 3 Results of voter mobilization experiments

Study	*Date*	*Election*	*Place*	*N of subjects (including control group)*	*Treatment*	*Effects on turnout*[*]
Gosnell (1927)	1924	Presidential	Chicago	3,969 registered voters	Mail	+1%
Gosnell (1927)	1925	Mayoral	Chicago	3,676 registered voters	Mail	+9%
Eldersveld (1956)	1953	Municipal	Ann Arbor	41 registered voters	Canvass	+42%
				43 registered voters	Mail	+26%
Eldersveld (1956)	1954	Municipal	Ann Arbor	276 registered voters	Canvass	+20%
				268 registered voters	Mail	+4%
				220 registered voters	Phone	+18%
Miller et al. (1981)	1980	Primary	Carbondale, IL	79 registered voters	Canvass	+21%
				80 registered voters	Mail	+19%
				81 registered voters	Phone	+15%
Adams and Smith (1980)	1979	Special city council	Washington, DC	2,650 registered voters	Phone	+9%
Gerber and Green (2000)	1998	General	New Haven, CT	29,380 registered voters	Canvass	+9%
				29,380 registered voters	Mail	+1%
				29,380 registered voters	Phone	−4%

[*]These are the effects reported in the tables of these research reports. They have not been adjusted for contact rates.
In Eldersveld's 1953 experiment, subjects were those who opposed or had no opinion about charter reform. In 1954, subjects were those who had voted in national but not local elections. Note that this table includes only studies that use random experimental design (or near-random, in the case of Gosnell [1927]). It excludes "controlled" experiments such as Blydenburgh (1971).

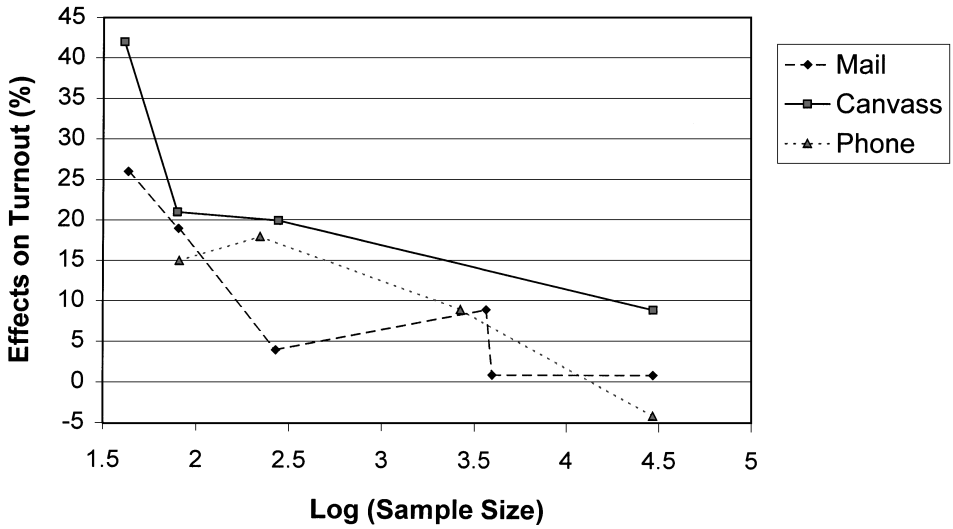

Fig. 1 Relationship between sample size and effect size.

to produce findings that overestimate the true effects, and published treatment effects will decrease as the sample size increases. The published literature is consistent with both of these predictions. Sending voters mailers prior to the election has immense effects when the sample size is small and relatively muted effects when the sample size is large. For example, while the classic Eldersveld (1956) study of the 1953 election found mail to have a massive, 26-percentage-point effect based on a sample size of 43, this figure dropped to 4 when the sample size was expanded to 268 the following year. Using the logged sample sizes of the mail experiments as predictors of the logged effects sizes, we find a slope of -0.46 (SE $= 0.15$). Similarly, the effects of phone calls and personal canvassing tend to be much larger in small studies than big ones. The effects of personal canvassing, for example, is found to be 9 percentage-points in Gerber and Green's (2000) study of 29,380 subjects, compared to 42 percentage-points in Eldersveld's study of 41 subjects prior to the 1953 election. Finally, the four phone experiments suggest that effect size declines as one moves from the small studies by Eldersveld (1956) and Miller et al. (1981) to those by Adams and Smith (1980) and Gerber and Green (2000).

The strong relationship between sample size and effect size is clearly shown in Fig. 1. For all modes of voter contact, studies with larger samples produce smaller estimated effects. Consistent with the examples presented earlier, when samples are small, all modes of treatment appear to work very well. While we cannot observe either the "true effect" of various modes of contact or the publication rule directly, it is nonetheless interesting to compare the pattern in the published studies to the examples shown in Table 1. Suppose that the true effects of mail, canvassing, and phone calls are similar to those reported in the largest study of voter contact, that by Gerber and Green (2000). For purposes of comparison to the cases analyzed in Table 1, let the true effect of canvassing be 10%, and let the true effect of mail and phone equal 1%.[5] In the absence of publication bias, we expect no tendency for smaller studies to show larger or smaller estimated effects. However, when the sample

[5]Setting the true effect of phone calls to 1% is a matter of convenience. In the presence of publication bias, until sample sizes reach into the thousands, the published literature when the true effect is 1% is very similar to the published literature when the effect is 0%.

size is 50 and the true treatment effect is 10%, Table 1 shows that the cutoff rule produces expected values of the published literature of 31%. When the true effect is only 1%, Table 1 shows that the predicted effects are 29%. Now consider the actual studies. The two smallest studies assess both mail and canvassing. The average sample size is slightly over 50, and the average treatment effects are 31% for canvassing and 23% for mailings, results consistent with the case of a fairly strict cutoff rule. A comparison of the published effects of phone calls reveals a similar pattern.

4 Conclusion

The pattern of results observed in the voter mobilization literature matches that predicted by the model of publication bias. Smaller studies indeed report larger effect sizes. While the literature on voter mobilization cannot be said to be typical of publications in political science or social science more generally, it provides interesting evidence of the existence of publication bias. Although nonexperimental research often relies on larger samples, nonexperimental literatures may be even more susceptible to publication bias. Uncertainties about model specification make data analysis much more complex and arguably dependent on the discretion of the analyst. Also, nonexperimental data are more abundant and accessible than experimental data, so it may be easier to locate or gather data that yield significant effect estimates. The method used in this paper could well be applied to nonexperimental literatures, gauging the correlation between sample size and effect size.

Publication bias poses a potentially serious problem to those who endeavor to synthesize research findings in political science. Unfortunately, for any given literature it is difficult to know how many other studies were conducted but went unreported. It would be helpful if researchers automatically published a synopsis of their findings regardless of the outcome. More realistically, professional associations within political science could create a central registry of abstracts for proposed studies, akin to what exists in the medical sciences (Chalmers et al. 1986; Simes 1986; Begg and Berlin 1988). Meta-analysts could then compare the published with the unpublished literature to get a more accurate sense of what was left on the cutting-room floor. If the discipline is to take seriously its large and growing research output, it must foster institutions such as registries that allow for meaningful syntheses of existing findings. Until then, political scientists are probably well advised to discount deeply the diagnostic weight of published studies based on small numbers of observations.

References

Adams, William C., and Dennis J. Smith. 1980. "Effects of Telephone Canvassing on Turnout and Preferences: A Field Experiment." *Public Opinion Quarterly* 44 (Autumn):389–395.

Begg, C. B. 1985. "A Measure to Aid in the Interpretation of Published Clinical Trials." *Statistics in Medicine* 4:1–9.

Begg, C. B. 1994. "Publication Bias." In *The Handbook of Research Synthesis*, eds. H. Cooper and L. V. Hedges. New York: Russell Sage Foundation.

Begg, C. B., and J. A. Berlin. 1988. "Publication Bias: A Problem in Interpreting Medical Data." *Journal of the Royal Statistical Society Series B* 151:419–463.

Berlin, J. A., C. B. Begg, and T. A. Louis. 1989. "An Assessment of Publication Bias Using a Sample of Published Clinical Trials." *Journal of the American Statistical Association* 84:381–392.

Chalmers, I., J. Hetherington, M. Newdick, L. Mutch, A. Grant, E. Enkin, and K. Dickersin. 1986. "The Oxford Database of Perinatal Trials: Developing a Register of Published Reports of Controlled Trials." *Controlled Clinical Trials* 7:306–324.

Coursol, A., and E. Wagner. 1986. "Effect of Positive Findings on Submission and Acceptance Rates: A Note on Meta-Analysis Bias." *Professional Psychology* 17:136–137.

DeLong, J. B., and K. Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100:1257–1272.

Dickersin, K. 1990. "The Existence of Publication Bias and Risk Factors for Its Occurrence." *Journal of the American Medical Association* 263:1385–1389.

Eldersveld, Samuel J. 1956. "Experimental Propaganda Techniques and Voting Behavior." *American Political Science Review* 50 (Mar.):154–165.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Direct Mail, and Telephone Contact on Voter Turnout: A Field Experiment." *American Political Science Review* 94 (Sept.):653–663.

Greene, William H. 1997. *Econometric Analysis*, 3rd ed. Upper Saddle River, NJ: Simon and Schuster.

Greenwald, A. G. 1975. "Consequences of Prejudice Against the Null Hypothesis." *Psychological Bulletin* 82:1–12.

Iyengar, S., and J. B. Greenhouse. 1988. "Selection Models and the File Drawer Problem." *Statistical Science* 3:9–135.

Light, R. J., and D. B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press.

Mahoney, M. J. 1977. "Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System." *Cognitive Therapy Research* 1:161–175.

Miller, Roy E., David A. Bositis, and Denise L. Baer. 1981. "Stimulating Voter Turnout in a Primary: Field Experiment with a Precinct Committeeman." *International Political Science Review* 2(4):445–460.

Sigelman, Lee. 1999. "Publication Bias Reconsidered." *Political Analysis* 8:201–210.

Simes, R. J. 1986. "Publication Bias: The Case for an International Registry of Clinical Trials." *Journal of Clinical Oncology* 4:1529–1541.