Quarterly Journal of Political Science, 2008, 3: 313–326

Research Note

Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals

Alan Gerber[1] and Neil Malhotra[2]

[1]*ISPS, Yale University, 77 Prospect Street, New Haven, CT 06520, USA; alan.gerber@yale.edu*
[2]*Graduate School of Business, Stanford University, Stanford, CA 94305-5015, USA; neilm@stanford.edu*

ABSTRACT

We examine the *APSR* and the *AJPS* for the presence of publication bias due to reliance on the 0.05 significance level. Our analysis employs a broad interpretation of publication bias, which we define as the outcome that occurs when, for whatever reason, publication practices lead to bias in the published parameter estimates. We examine the effect of the 0.05 significance level on the pattern of published findings using a "caliper" test, a novel method for comparing studies with heterogeneous effects, and find that we can reject the hypothesis of no publication bias at the 1 in 32 billion level. Our findings therefore raise the possibility that the results reported in the leading political science journals may be misleading due to publication bias. We also discuss some of the reasons for publication bias and propose reforms to reduce its impact on research.

A key objective of political science research is the accurate measurement of causal effects. Methodological advances (such as increased use of natural, laboratory, and field experiments) have made it much more plausible than in earlier decades that the results of individual studies are unbiased estimates. Unfortunately, better research design does not ensure unbiased literatures. For instance, if some results are more likely to be published, then literatures will be biased even if each study is done well.

Supplementary Material available from:
http://dx.doi.org/10.1561/100.00008024_supp
MS submitted 3 October 2007; *final version received 8 April 2008*
ISSN 1554-0626; DOI 10.1561/100.00008024

This paper examines the publication performance of leading journals to see if there is evidence of publication bias. We adopt a broad interpretation of publication bias, which we define as the outcome that occurs when, for whatever reason, publication practices lead to bias in the published parameter estimates. This can happen in several ways: (1) editors and reviewers may prefer significant results and reject methodologically sound articles that do not achieve certain statistical significance thresholds; (2) scholars may only submit studies with statistically significant results to journals and place the rest in "file drawers"; (3) investigators may adjust sample sizes after observing that results narrowly fail tests of significance; and (4) researchers may engage in data mining to find model specifications and sub-samples that achieve significance thresholds.[1] All of these mechanisms produce pooled estimates that are biased, in that they are not collectively equal to the true population parameter. Although disentangling the sources of bias would be interesting, this paper focuses on the prior task of evaluating what is published in the leading journals to assess whether the published results do, in fact, display telltale signs of bias.

Publication bias has been found to be ubiquitous in several academic disciplines, including medical research (e.g., Begg and Berlin 1988, Berlin *et al.* 1989, Easterbrook *et al.* 1991, Levois and Layard 1995), psychology (e.g., Sterling 1959, Greenwald 1975, Coursol and Wagner 1986), economics (e.g., De Long and Lang 1992, Card and Krueger 1995, Ashenfelter *et al.* 1999, Doucoulagos 2006), and sociology (Gerber and Malhotra 2008). A more detailed literature review can be found in Gerber and Malhotra (2008). In political science, investigations of publication bias are virtually nonexistent. After Sigelman (1999) raised publication bias as a potential concern in the discipline, Gerber *et al.* (2000) analyzed the voting mobilization literature and found a strong, negative relationship between effect size and sample size, a pattern consistent with publication bias based on rejection or nonsubmission of statistically insignificant findings. Considering that extensive evidence of publication bias has been found throughout the natural and social sciences since the 1950s, it is surprising that there is no extensive analysis of political science research.

Standard methods for detecting publication bias assess the pattern of results reported across the set of published studies measuring the effects of a particular treatment, often in the context of meta-analysis. One sign that publication bias based on the 0.05 significance level is present in a literature is if smaller studies tend to report larger results, which is assessed statistically via a funnel graph (Light and Pillemar 1984). The association of sample size with effect size suggests publication bias since for a small-N study not finding a large estimated treatment effect, a result which may be produced by chance, the ratio of the estimated treatment effect to the standard error will not exceed the critical values needed to generate a statistically significant t-ratio. Consequently, the paper may be less likely to be submitted or accepted. Alternatively, after observing the study falls short of significance, the researcher may collect more data and then publish only if the

[1] If these models are incorrectly specified, then these published studies will be biased, meaning that publication bias in the broader literature is an aggregation of bias in individual studies. Hence, results may not be stored in file drawers, but "tweaked" until statistical significance is achieved.

significance threshold is broken, a practice which would produce a biased literature. A different method must be used to detect publication bias across studies which each examine a different subject matter, since researchers anticipating larger treatment effects might reasonably plan to use a smaller sample. In other cases, sample sizes are effectively fixed (e.g., cross-sections of OECD countries or US states). Moreover, a funnel graph requires that studies employ common dependent variables and comparable measures of effect size.

In this paper, we examine publication bias by considering all statistical studies on all topics published in the *American Political Science Review* (*APSR*) and the *American Journal of Political Science* (*AJPS*) over the past 13 years. We develop a simple and intuitive test for the presence of publication bias. We examine the ratio of reported results just above and just below the critical value associated with the 0.05 p-value, a test we will refer to as a "caliper test," since it is based on the rate of reported occurrence of test statistics within a narrow band. We find that there are far more reported coefficients just above the critical value than just below it. The probability that some of the patterns we find are due to chance is less than 1 in 32 billion. Our findings raise the possibility that the results reported in the leading political science journals may be misleading due to publication bias.

METHOD AND DATA

To examine a diverse group of studies with heterogeneous dependent variables, we employ a novel method which we term the "caliper test." The test detects publication bias by comparing the number of observations in equal-size intervals just below and just above the threshold value for statistical significance.[2] If the imbalance is sufficiently great, the probability this imbalance is due to chance is small and the null hypothesis that the collection of results reported in the journals is not affected by critical values is rejected. A formal presentation of the "caliper test" is provided in Gerber and Malhotra (2008). The intuition supporting the test is straightforward. Since the sampling distribution that generates a coefficient estimate can be viewed as a continuous probability distribution, the chance of observing a result just above any arbitrary number should be about the same as the chance of observing a result just below it. Of course, the caliper test only tests for publication bias in a local neighborhood around the critical values, and cannot

[2] After our research was concluded we became aware of a manuscript by Edward Tufte, "Evidence Selection in Statistical Studies of Political Economy: The Distribution of Published Statistics," which finds substantial evidence of publication bias. Among other things, Tufte examines the distribution of reported results for t-tests around the critical value of 2. He reports that, for studies dealing with macroeconomic fluctuations and the national vote, t-values in the interval 2.0 to 2.4 were three times more likely than values between 1.6 and 2.0. He also looks at a second literature, the effect of macroeconomic fluctuations on presidential popularity, to confirm this finding out of sample. He finds a much lower imbalance; values in the interval 2.0 to 2.4 are 50% more likely than values between 1.6 and 2.0. Given the subject of this paper, it is ironic that Tufte's manuscript (circa 1985, the date written in ink on our copy) appears to be both unpublished and never cited.

evaluate bias in other areas of the distribution, which may take other forms. As discussed below, in the later stages of a literature, there may exist incentives to publish contrarian findings with z-scores closer to zero.

Our dataset was extracted from the studies reported over a thirteen-year period in the *APSR* and the *AJPS*, two of the leading journals in the discipline. These journals are of obvious interest because they have published a huge share of the discipline's most highly regarded and influential scholarship and remain prestigious outlets for new work. We also examined the *QJPS*, but there are too few issues and articles per issue for meaningful analysis.[3] When performing an assessment of a single parameter, such as a treatment effect in a medical study or an elasticity of labor supply, it is relatively easy to extract the relevant data from the publications, and then construct a funnel graph. In our investigation, which aims to analyze where the coefficient relevant for the author's hypothesis test falls relative to the critical value for the hypothesis test, we must first determine what the hypotheses are and then find the coefficients related to the hypotheses. This requires a selection methodology and in this section we will briefly describe the procedures we used. A more detailed discussion can be found in the appendix.

In conducting the caliper tests, we analyzed volumes 89–101 (1995–2007) from the *APSR* and volumes 39–51 (1995–2007) from the *AJPS*, and selected the empirical papers. To avoid difficult judgments about authorial intentions, we only examined papers that listed a set of hypotheses prior to presenting the statistical results.[4] Papers use slightly different conventions for naming hypotheses in this fashion (e.g., "Hypothesis 1," "H1," "H_1," "The first hypothesis").[5] We further restricted our sample to those articles that had 35 or fewer coefficients linked to the hypotheses, a procedure we justify in greater detail in the appendix. This appeared to be a natural cutoff as the article with the next fewest coefficients had 45.

The next step is to link the hypotheses to specific regression coefficients. This could usually be done by inspecting regression tables, which often stated the hypotheses associated with each coefficient.[6] We confirmed that we were recording the correct figures by reading sections of the paper identifying which regression coefficients applied to which hypotheses. We remained agnostic about which regressions to include in the dataset,

---

[3] Using our selection methodology, only two articles were extracted from the two complete volumes of the *QJPS*. Only six coefficients were contained within the 10% caliper — four were above and two were below.

[4] An alternative approach would have been to have multiple coders and a method for resolving disputes over what the author wished to test and how it related to the reported results.

[5] The vast majority of these hypotheses required a statistically significant result to be accepted. Only three articles (out of 137 in the sample) had a hypothesis predicting a null finding. These articles constitute ten coefficients (out of 1811), all of which are statistically insignificant but none of which are in any caliper tested. Hence, these "null hypotheses" do not affect the results presented below.

[6] In some cases, the reported t-statistic (estimate divided by standard error) was inconsistent with the notation of significant (i.e., presence of a "star"), due to rounding error. In these cases, we were conservative and assumed that the precise t-statistic was below the caliper if significance was not indicated, but above the caliper if significance was noted.

including all specifications (full and partial) in the data analysis. We present an example of how regression coefficients were selected in the appendix.

RESULTS

Figures 1(b)(a) and 1(b)(b) show the distribution of z-scores[7] for coefficients reported in the *APSR* and the *AJPS* for one- and two-tailed tests, respectively.[8] The dashed line represents the critical value for the canonical 5% test of statistical significance. There is a clear pattern in these figures. Turning first to the two-tailed tests, there is a dramatic spike in the number of z-scores in the *APSR* and *AJPS* just over the critical value of 1.96 (see Figure 1(b)(a)). The formation in the neighborhood of the critical value resembles

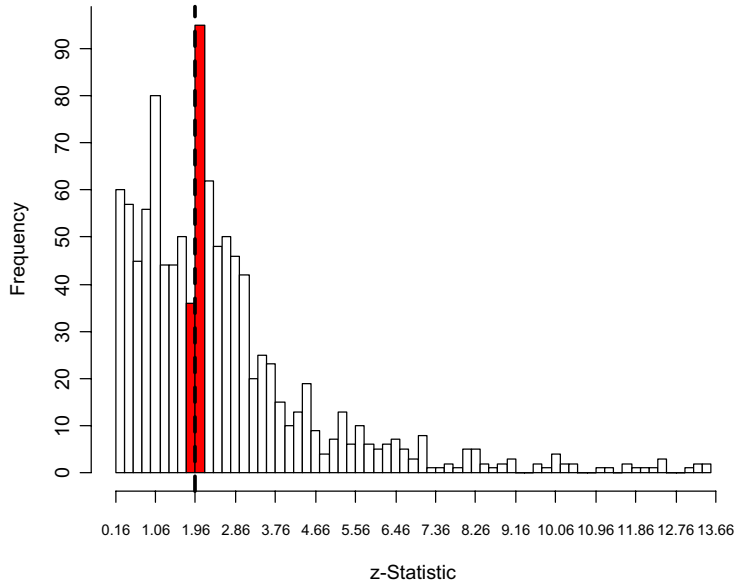

Figure 1(a). Histogram of z-statistics, *APSR & AJPS* (Two-Tailed). Width of bars (0.20) approximately represents 10% caliper. Dotted line represents critical z-statistic (1.96) associated with $p = 0.05$ significance level for one-tailed tests.

---

[7] The formal derivation of the caliper test is based on z-scores. However, we replicated the analyses using t-statistics, and unsurprisingly, the results were nearly identical. Generally, studies employed sufficiently large samples, and there were very few coefficients in the extremely narrow caliper between 1.96 and 1.99.

[8] Very large outlier z-scores are omitted to make the x-axis labels readable. The omitted cases are a very small percentage (between 2.4% and 3.3%) of the sample and do not affect the caliper tests. Additionally, authors make it clear in tables whether they are testing one-sided or two-sided hypotheses.

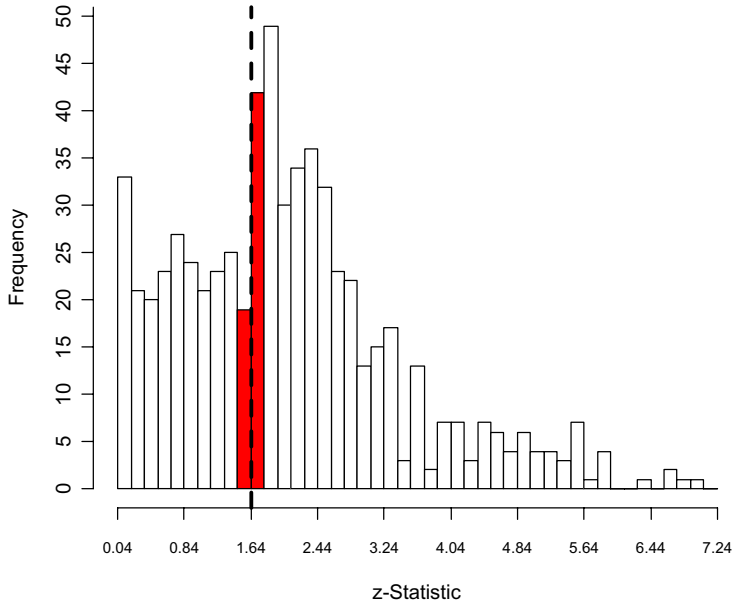

Figure 1(b). Histogram of z-statistics, *APSR & AJPS* (One-Tailed). Width of bars (0.16) approximately represents 10% caliper. Dotted line represents critical z-statistic (1.64) associated with $p = 0.05$ significance level for one-tailed tests.

a steep cliff that drops into a valley, suggesting that a portion of the density below the critical value has been cut out and deposited above the critical value. Turning to the one-tailed tests, we again see the spike, which this time appears just over the critical value of 1.64 (see Figure 1(b)(b)). These distributions show that there is greater density of published findings in the region barely above the arbitrary 0.05 significance level compared to the region barely below it. One notable feature is that the number of cases in the interval just over the critical value is the *global maximum* in Figure 1(b)(a) and the second highest number in Figure 1(b)(b), whereas the interval just below the critical value is the local minimum in both cases.

Figures 1(b)(a) and 1(b)(b) use a 10% caliper. Table 1 shows the results using a variety of alternative calipers, with one- and two-sided tests pooled. Panel A of Table 1 shows the results for the *APSR* and Panel B shows the results for the *AJPS*. Each panel shows the results for the entire thirteen-year sample as well as the results broken into the first seven years and the next six years. The results depicted in the figures are robust across journals, caliper sizes, and publication dates.[9] However, there is a pattern in the results.

[9] We also analyzed the data by journal, pooling one- and two-tailed tests. Although the pattern is still quite pronounced, the effect is somewhat dampened, suggesting the presence of strategic selection of one-sided versus two-sided hypothesis testing. For the *APSR* (using a 10% caliper), the ratio of

Table 1. Caliper tests of publication bias in *APSR* and *AJPS*

	Over Caliper	Under Caliper	p-value*
A. APSR			
Vol. 89–101			
10% Caliper	49	15	< 0.001
15% Caliper	67	23	< 0.001
20% Caliper	83	33	< 0.001
Vol. 96–101			
10% Caliper	36	11	< 0.001
15% Caliper	46	17	< 0.001
20% Caliper	55	21	< 0.001
Vol. 89–95			
10% Caliper	13	4	0.02
15% Caliper	21	6	0.003
20% Caliper	28	12	0.008
B. AJPS			
Vol. 39–51			
10% Caliper	90	38	< 0.001
15% Caliper	128	66	< 0.001
20% Caliper	165	95	< 0.001
Vol. 46–51			
10% Caliper	56	25	< 0.001
15% Caliper	80	45	0.001
20% Caliper	105	66	0.002
Vol. 39–45			
10% Caliper	34	13	0.002
15% Caliper	48	21	< 0.001
20% Caliper	60	29	< 0.001

*Based on density of binomial distribution (one-tailed).
Note: "Over Caliper" indicates number of results that are between 0%–X% greater than critical value (1.64 and 1.96 for one- and two-tailed tests, respectively), where X is the size of the caliper. For instance, for the 10% caliper, the "Over Caliper" range is approximately 1.64–1.81 for one-tailed tests and 1.96–2.16 for two-tailed tests. "Under Caliper" represents the number of results that are between 0%–X% less than the critical value. For the 10% caliper, the "Under Caliper" range is about 1.48–1.64 (1.76–1.96).

The ratio of findings just below to just over the critical values decreases as the caliper size expands, which is consistent with the idea that there is something unusual about the critical value. The marginal distribution of findings begins to approximate the uniform more closely in the outer rungs of the caliper. For instance, whereas the ratio is over 3:1 for the *APSR* for the 10% caliper, the ratio is only 2.5:1 for the 20% caliper. The fact that the disparities are most dramatic in the narrow regions nearest to the critical value provides additional evidence the imbalance is from publication bias rather than a chance occurrence.

Under the null hypothesis that a z-score just above and just below an arbitrary value is about equally likely, and that the coefficients are statistically independent, we calculated the chance that imbalances of the magnitude we observe would occur by chance. For the 10% caliper for the *APSR*, the chance that we would observe an imbalance as large or larger than we do by chance alone is less than 1 in 80,000. For the *AJPS*, the chance is less than 1 in 400,000. The joint probability is astronomically small, approximately 1 in 32 billion.[10]

The statistical tests we perform may overstate the rarity of the patterns we report. One factor that complicates the analysis is that the studies included in our dataset often contributed more than one coefficient. This suggests each coefficient cannot be viewed as statistically independent, though the departure from independence over the narrow range of values included in the caliper is almost certainly trivial. Nevertheless, we have no way to determine the covariances from the information contained in the articles. One further check of the results is to restrict attention to only those studies that contributed one coefficient to the statistical test. Table 2 shows the results for the *APSR* and the *AJPS* broken down by the number of coefficients per study that fell within the range associated with various caliper values. Restricting attention to only studies that contributed a single coefficient shows the same proportional imbalance reported earlier. The odds of an imbalance as great or greater than the 35:12 imbalance found for the *APSR* and *AJPS* combined using the 10% caliper is, under the hypothesis of equal probability, less than 0.0006.

It is possible that the *APSR* and *AJPS* specialize in statistically significant results, while lesser journals publish the remaining studies. If so, the problem of publication bias would be mitigated since all the studies would be published somewhere, just not in the top journals. To determine whether the results we find for the two journals studied here also apply to other outlets, we examined the published literature appearing in a broader collection of journals on two topics with large bodies of work: the effect of negative advertising and economic voting. In a separate paper, we demonstrate similar patterns of imbalance via a caliper test (Gerber and Malhotra 2006).

---

findings above and below the caliper is 17:8 for the 1.64 critical value and 34:20 for the 1.96 critical value. For the *AJPS*, the ratio of findings is 38:40 and 72:46 for the 1.64 and 1.96 critical values, respectively.

[10] We do not interpret the magnitude of the results in terms of "effect sizes," because we do not know the true parameter estimates and hence any calculations would be highly speculative. Accordingly, any statements about the size or extent of bias could be misinterpreted as identifying a particular source of the bias.

Table 2. Caliper tests of *APSR* and *AJPS* by coefficients contributed per study

Coef. Contributed	10% Caliper			15% Caliper			20% Caliper		
	Over	Under	Studies	Over	Under	Studies	Over	Under	Studies
A. APSR									
1	9	3	12	8	2	10	8	2	10
2	9	1	5	10	2	6	5	3	4
3	13	5	6	13	2	5	15	6	7
4	8	0	2	2	2	1	8	4	3
5	6	4	2	12	3	3	10	0	2
6	4	2	1	7	5	2	3	3	1
7	0	0	0	10	4	2	8	6	2
8	0	0	0	5	3	1	12	4	2
9	0	0	0	0	0	0	8	1	1
10	0	0	0	0	0	0	6	4	1
Total	49	15	26	67	23	30	83	33	33
B. AJPS									
1	25	10	35	16	6	22	14	5	19
2	18	10	14	26	16	21	19	13	16
3	9	3	4	20	19	13	30	27	19
4	26	10	9	24	12	9	21	11	8
5	4	1	1	4	1	1	21	14	7
6	8	4	2	28	8	6	17	7	4
7	0	0	0	10	4	2	21	7	4
8	0	0	0	0	0	0	17	7	3
9	0	0	0	0	0	0	5	4	1
Total	90	38	65	128	66	74	165	95	81

Note: Coefficients contributed refers to the number of results for a given study within a certain caliper. For instance, for the *APSR*, one study had six coefficients within the 10% caliper — four over it and two below it.

DISCUSSION

Our results provide empirical verification that as reviewers, editors, and researchers, political scientists appear to be quite conscious of the 0.05 significance level. As we outline earlier in the paper, publication bias may be produced by several factors — (1) journal decisions; (2) submission practices of scholars; and (3) research practices of scholars, whose behavior is based in part on expectations about how journals evaluate research. With access to a full set of submissions — not only those that are published — future

research could determine the extent to which bias is due to selection on the part of reviewers and editors.[11]

Are there any other, more innocuous explanations for the imbalance we observe? It has been argued that studies with statistically significant findings are often better studies. From this it follows that one would expect to see a surplus of such studies, especially in top journals. However, this is not sufficient to explain our findings, since it is impossible to credit the claim that the quality of the research methodology deteriorated in a discontinuous fashion just below and improved in a discontinuous fashion just above the z-score associated with the 5% significance level.

There is convincing evidence that something is causing a distortion of the published literature around the critical values for statistical significance. How consequential is this? If the extent of the damage to science is a small distortion in the reported results, then publication bias is an unfortunate source of error but perhaps not catastrophic. This somewhat sanguine view is probably not justified. First, publication bias may result in a significant understatement of the chances of a Type I error, which lends false confidence and may misdirect subsequent research. Second, anticipation of journal practices may distort how studies are conducted, encouraging data mining, specification searches, and *post hoc* sample size adjustments. Third, and perhaps most important, holding work to the arbitrary standard of $p < 0.05$ may discourage the pursuit and publication of work that is well designed and on important topics but unlikely to produce precisely measured estimates. While statistical insignificance may suggest poor design and theorizing, certainly the $p < 0.05$ standard should not be used as a proxy for quality research by expert scholars. An excellent research design that produces imprecise though unbiased measurements can be extremely valuable. A collection of well identified though imprecise studies (each with statistically insignificant results) can, together, be far more valuable than large studies with more precise, though biased measurement (Gerber *et al.* 2004).[12] Finally, we have focused on only one symptom of publication bias, excessive occurrence of z-scores just over 2. Our discussions assume that distortions would always push z-scores over the critical value. In fact, in some cases there might be a greater interest and professional reward to uncovering a statistically insignificant effect (for example, in the later stages of a literature). This would suggest that the amount of distortion of estimates due to the critical values captured by our analysis is only a portion of the total.

Publication bias is an ancient problem and suggestions to address it are speculative.[13] One institutional reform that might limit the number of "lost" studies and specification/subgroup/hypothesis searching is the establishment of study registries. Prior to conducting research, a description of the proposed research would be filed with a central

---

[11] We attempted to conduct such an analysis; however, our request to the editor of the *APSR* for a sample of submissions was denied on privacy grounds, and we did not receive a response from the editor of the *AJPS*.

[12] Gerber *et al.* (2004) compare the value of alternative research designs as a function of the degree of uncertainty regarding the bias in the estimates produced by the alternative methods, as well as the amount and types of prior research.

[13] In a letter to the editor of the *Lancet*, Mark Pettricrew (1998) mentioned that publication bias was raised as a concern by Diagoras, the fifth century Greek poet.

registry, along with the proposed sample size. In medicine, some journals now refuse to publish studies that failed to file a description of the proposed research design and analysis with a registry prior to performing the work. While the use of registries has become common for experimental research and should be considered for political science experimentation as well, we propose that registries can be usefully extended to the class of observational studies where the hypotheses to be tested are formulated in advance and the data are not yet available to the researcher, such as research using survey or election results.

Echoing Gill's (1999) critique of the logic of null hypothesis significance testing, it would appear sensible for authors and reviewers to place greater emphasis in their write-ups and evaluations on confidence intervals and the substantive importance of the effects rather than whether or not a coefficient estimate achieves the $p < 0.05$ standard. More generally, while the estimates in a study under review will affect the publication decision, publication bias might be avoided if greater attention was given to the research design rather than the estimates produced by a particular study. Since an innovation in research design can improve all subsequent work, there should be strong encouragement for innovative studies that show how a causal effect or important quantity can be estimated in a convincing fashion. For such studies, it may sometimes be the case that data is scarce or measurement difficult and, consequently, the power of the particular study is weak. Nevertheless, since a strong design can point the way to subsequent innovations, there are good reasons to publish the work even if the results fall short of some ancient guidelines conventionally employed in evaluating hypotheses.

APPENDIX: PROCEDURES FOR COLLECTING STUDIES FROM THE *APSR* AND THE *AJPS*

In this appendix, we detail how coefficients were selected. In conducting the caliper tests, we analyzed volumes 89–101 (1995–2007) from the *APSR* and volumes 39–51 (1995–2007) from the *AJPS*. The total number of articles and research notes (excluding forums, workshops, book reviews, controversies, and presidential addresses) yielded 555 papers from the *APSR* and 737 from the *AJPS*. Not all 1,292 papers could be tested for publication bias using the caliper method. Political theory papers and pure formal models did not contain statistical tests. For those papers where there was statistical analysis, we attempted to formulate a replicable procedure to extract coefficients and standard errors. To avoid difficult judgments about authorial intentions, we eliminated articles that ran regressions but did not explicitly state hypotheses.

We only examined papers that listed a set of hypotheses prior to presenting the statistical results. Papers use slightly different conventions for naming hypotheses in this fashion (e.g., "Hypothesis 1," "H1," "H_1," "The first hypothesis"). Further, some authors italicize or bold hypotheses within paragraphs while others indent them. Of the original 1,292 articles collected, 219 employed this convention of explicitly listing hypotheses (76 in the *APSR* and 143 in the *AJPS*). We further restricted our sample to those articles that had 35 or fewer coefficients linked to the hypotheses. This appeared to be a natural cutoff as the article with the next fewest coefficients had 45. Applying

this rule, we excluded 82 articles (38 from the *APSR* and 44 from the *AJPS*). There are two rationales for this last exclusion. First, it minimizes the influence of any one article. Second, it is unclear what publication bias hypotheses predict for a paper with many coefficients. We suspect that publication bias is related to the important results, and including these articles would require judgment on our part as to which estimates were the most "important." Conversely, we do not need to make such decisions with articles that reported relatively few results. Hence, we are left with 137 articles (38 from the *APSR* and 99 from the *AJPS*) of the original 1,292 to subject to the caliper test. These selection steps are summarized below:

	APSR	AJPS	Total
Total articles	555	737	1292
Total that list Hypotheses	76	143	219
Selected articles ($\leq$ coeff.)	38	99	137

We now provide an example of a selected study, and how coefficients were linked to specific hypotheses. Hoekstra (2000, *APSR*) states the following four hypotheses at the beginning of the paper:

HYPOTHESIS 1a. *Residents in and around the immediate community in which a case originates should exhibit a higher level of awareness of the Court's decision than is usually found in national samples.*

HYPOTHESIS 1b. *Those within the immediate community should show a high level of awareness than those who reside in the surrounding areas.*

HYPOTHESIS 2. *Among those who hear of the Court's decision, those from the immediate community should find the case more important than those from the surrounding areas.*

HYPOTHESIS 3. *Awareness of the Court's decision should affect evaluations of the Supreme Court. Those who initially approve/disapprove of the policy position articulated in the decision should show increased/decreased support for the Court.*

HYPOTHESIS 4. *The effect of hearing about the Court's decision should be greatest among those who reside in the immediate community. They will show greater change in evolution of the Court following its decision, above and beyond that predicted for those who reside outside the community.*

For the articles that were selected, we read the text of the article to associate each hypothesis with regression coefficients presented in the tables. To remain agnostic, we included all coefficients associated with the hypothesis across multiple specifications. Here is an example of how regression coefficients were selected using the Hoekstra (2000) article:

HYPOTHESIS 1a. *The paper does not test this hypothesis directly because the author does not have national data. She just uses a content analysis* (*no statistical tests*).

HYPOTHESIS 1b. *All hypotheses are tested for four different Court cases in four different geographical areas in Table 3. The variable of interest is "Town of residence," for which there are 4 coefficients* (*one for each area*).

HYPOTHESIS 2. *Hypotheses are tested with difference-of-means tests in Table 5. From the reported p-values, we extracted the z-statistics. For each area, we specifically looked at the summary measure of importance, as the text suggested. For Center Moriches, this statistic was not measured so we recorded the only one available: "strength of opinion." Thus, we recorded an additional four coefficients* (*one from each area*).

HYPOTHESES 3 and 4. *These hypotheses are tested in Table 6 for three of the geographic regions using an interaction term between "Initial opinion" and "Town of residence." Hence, three additional coefficients were recorded.*

Therefore, in total, 11 coefficients were recorded from Hoekstra (2000).

REFERENCES

Ashenfelter, O., C. Harmon, and H. Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias." *Labour Economics* 6(4): 453–470.

Begg, C. B., and J. A. Berlin. 1988. "Publication Bias: A Problem in Interpreting Medical Data." *Journal of the Royal Statistical Society Series A* 151(3): 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(406): 381–392.

Card, D., and A. B. Krueger. 1995. "Time-Series Minimum Wage Studies: A Meta- Analysis." *American Economic Review* 85(2): 238–243.

Coursol, A., and E. E. Wagner. 1986. "Effect of Positive Findings on Submission and Acceptance Rates: A Note on Meta-Analysis Bias." *Professional Psychology, Research and Practice* 17(2): 136–137.

Doucouliagos, C. 2005. "Publication Bias in the Economic Freedom and Economic Growth Literature." *Journal of Economic Surveys* 19(2): 367–387.

De Long, J. B., and K. Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100(6): 1257–1272.

Easterbrook, P. J., J. C. Keruly, T. Creagh-Kirk, D. D. Richman, R. E. Chaison, and R. D. Moore. 1991. "Racial and Ethnic Differences in Outcome in Zidovudine-Treated Patients with Advanced HIV Disease." *Journal of the American Medical Association* 19(266): 2713–2718.

Gerber, A. S., D. P. Green, and D. Nickerson. 2000. "Testing for Publication Bias in Political Science." *Political Analysis* 9(4): 385–392.

Gerber, A. S., D. P. Green, and E. H. Kaplan. 2004. "The Illusion of Learning from Observational Research." In *Problems and Methods in the Study of Politics*, eds. I. Shapiro, R. M. Smith, and T. E. Masoud, Cambridge, UK: Cambridge University Press, pp. 251–273.

Gerber, A. S., and N. Malhotra. 2006. "Can Political Science Literatures Be Believed? A Study of Publication Bias in the *APSR* and the *AJPS*." Society for Political Methodology Working Paper.

Gerber, A. S., and N. Malhotra. 2008. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods and Research* 37(1): 3–30.

Gill, J. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3): 647–674.

Greenwald, A. G. 1975. "Consequences of Prejudice Against the Null Hypothesis." *Psychological Bulletin* 82(1): 1–20.

Levois, M. E., and M. W. Layard. 1995. "Publication Bias in the Environmental Tobacco Smoke/Coronary Heart Disease Epidemiologic Literature." *Regulatory Toxicology and Pharmacology* 21(1): 184–191.

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

Petticrew, M. 1998. "Diagoras of Melos (500 BC): An Early Analyst of Publication Bias." *The Lancet* 352(9139): 1558.

Sigelman, L. 1999. "Publication Bias Reconsidered." *Political Analysis* 8(2): 201–210.
Sterling, T. D. 1959. "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa." *Journal of the American Statistical Association* 54(285): 30–34.
Tufte, E. R. 1985. "Evidence Selection in Statistical Studies of Political Economy: The Distribution of Published Statistics." Unpublished manuscript, Yale University.