

Publication Bias in Two Political Behavior Literatures

Alan S. Gerber
Yale University
Department of Political Science
Institution for Social and Policy Studies
77 Prospect Street, PO Box 208209
New Haven, CT 06520-8209
alan.gerber@yale.edu
203.432.5232 (voice), 203.432.3296 (fax)

Neil Malhotra
Stanford University
Stanford Graduate School of Business
518 Memorial Way
Stanford, CA 94305-5015
neilm@stanford.edu
408.772.7969 (voice)

Conor M. Dowling
Yale University
Institution for Social and Policy Studies
77 Prospect Street, PO Box 208209
New Haven, CT 06520-8209
conor.dowling@yale.edu
203.432.4811 (voice), 203.432.3296 (fax)

David Doherty
Yale University
Institution for Social and Policy Studies
77 Prospect Street, PO Box 208209
New Haven, CT 06520-8209
david.doherty@yale.edu
203.432.5057 (voice), 203.432.3296 (fax)

ABSTRACT

Publication bias occurs when the probability a paper enters the scholarly literature is a function of the magnitude or significance levels of the coefficient estimates. We investigate publication bias in two large literatures in political behavior, economic voting and the effects of negative advertising. We find that the pattern of published estimates is consistent with the presence of publication bias, and that bias is more prevalent in the most influential and highly cited outlets. We consider the possible causes and find some evidence that papers systematically employ one-sided hypothesis tests in response to failure to meet the more demanding critical values associated with two-tailed tests, a practice that leads to misleading reports of the probability of Type I errors.

Keywords: publication bias; political behavior; economic voting; negative advertising

Alan S. Gerber is the Charles C. and Doratheia S. Dilley Professor of political science and director of the Center for the Study of American Politics at Yale University. His current research focuses on the study of campaign communications, and he has designed experimental evaluations of many partisan and nonpartisan campaigns and fundraising programs. His research has appeared in the *American Political Science Review*, the *American Journal of Political Science*, the *Journal of Politics*, and the *Proceedings of the National Academy of Science*.

Neil Malhotra is Assistant Professor of Political Economy at the Stanford Graduate School of Business. His research interests include political behavior, legislative institutions, and survey methodology. His research has been published in the *American Political Science Review* and the *Journal of Politics*, among other outlets.

Conor M. Dowling is a postdoctoral associate at the Institution for Social and Policy Studies and the Center for the Study of American Politics at Yale University. His current research interests include campaigns and elections, political behavior, and research methodology. His work has appeared in *Political Analysis*, *Political Research Quarterly*, and *State Politics & Policy Quarterly*.

David Doherty is a postdoctoral associate in the Institution for Social and Policy Studies and the Center for the Study of American Politics at Yale University. He received his PhD in political science from the University of Colorado. His research interests are in public opinion, representation, and political psychology, and focus on how people interpret and evaluate the behavior of institutional actors.

Authors' Note: We acknowledge Andrew Gelman, Donald Green, Alexander Tahk, Jowei Chen, Alexander Kuo, Sarah Anzia, Jeremy Freese, Sean Riordan, Kendra Bischoff, four anonymous referees, and the editor for valuable suggestions. We also thank Ray Selie and Tania Juarez for helpful research assistance. Please address correspondence to Neil Malhotra, Stanford University, Stanford Graduate School of Business, 518 Memorial Way, Stanford, CA 94305-5015; e-mail: neilm@stanford.edu.

A large proportion of findings reported in political behavior are based on statistical analysis. If published work accurately represents the full body of research being conducted in an area and the reported hypothesis tests are constructed ex-ante, then researchers can be confident in their ability to interpret the magnitude of effects and the likelihood that they are due to chance. However, if the publication process is in some way biased, published work may present a distorted picture. Bias may enter at many points in the journey from analysis to publication (or failure to publish). If journal editors and reviewers tend to accept papers that include statistically significant findings—or researchers anticipate such a tendency—this may lead to the submission and publication of results that are the product of sampling error, fragile model specifications, or ex-post hypotheses.¹ Similarly, if studies that do not yield statistically significant results are never published—whether because they are never submitted or because they are rejected by reviewers—then those who read the published findings may erroneously assume that research questions are definitively answered when this is not the case. In sum, when the probability that a paper enters the scholarly literature is a function of the reported results or significance levels, researchers hoping to build on or refine previous findings may be led astray. More generally, publication bias is a small piece of the much larger question of how academic work, like all types of work, is shaped by professional incentives.

This is not the first study to investigate the prevalence of publication bias in social science research. A number of studies have identified publication bias in the fields of psychology (e.g., Sterling 1959; Greenwald 1975; Coursol and Wagner 1986), public health and medicine (e.g. Gotzsche 2006), economics (e.g., De Long and Lang 1992; Card and Krueger 1995; Ashenfelter et al. 1999; Doucouliagos 2005; Doucouliagos et al. 2005), and sociology (Gerber and Malhotra 2008a). However, relatively little work has been done to assess the degree of

publication bias in political science. Gerber and Malhotra (2008b) examine publication bias in the leading journals in political science. They find that in the *American Political Science Review* and *American Journal of Political Science* there are far more results just above critical values than can be explained by chance, a pattern which suggests that what is published in two of the top journals and how it is interpreted is influenced by arbitrary critical values of the t -distribution. It is conceivable that statistically significant findings are only published disproportionately in the most prestigious journals. If we look beyond these top journals this pattern may disappear. In this case the results of a study affect *where* the paper is published but not *whether* the paper is ultimately published.

To explore this possibility we examine two major literatures in political behavior—research on economic voting and research on the effects of negative advertising—to see if there is evidence of publication bias. A number of processes, including how editors and reviewers evaluate submitted research, how researchers decide what research to submit for review, and how researchers report their statistical tests, might explain any publication bias we observe. Although investigating the causes of publication bias is an important task, in this article we remain essentially agnostic about the extent to which various factors affect which studies are and are not published. Instead we focus on examining whether published parameter estimates in these two literatures indicate bias in the publication process.

We build upon and extend Gerber and Malhotra's analysis (2008b) in three ways. First, whereas they examined articles on all topics, we explore how publication bias influences findings in two specific literatures. Second, Gerber and Malhotra only analyzed studies published in two of the most prominent journals. By considering articles published across a wider set of journals, we can assess whether publication bias is more prevalent in more influential outlets.

Third, we provide novel evidence on one particular manifestation of publication bias: the strategic selection of one-tailed and two-tailed hypothesis tests based on the critical value.

The article is organized as follows. Section 1 provides an overview of our methodological approach. Section 2 describes how we constructed the data set for our statistical analysis. Section 3 presents the results for published studies in two bodies of literature in political science (the effects of negative advertising and studies on economic voting) that appear in a broad set of the discipline's journals. Section 4 discusses the implications of our findings.

Section 1. Methodological Overview

We examine publication bias by considering all statistical studies in two major literatures in political behavior published in ten top journals between 1990 and 2007. In the past, a number of approaches have been used to identify publication bias. Gerber, Green, and Nickerson (2000) found that the smaller the sample size used in published experimental voter mobilization studies, the larger the magnitude of the reported effects. The authors interpret this relationship as indicative of publication bias. Given that these studies all relied on an experimental design using similar treatments, there is little reason to expect a strong, negative relationship between sample size and effect size. Published studies with small sample sizes may exhibit larger effects because that is the only way that they can cross thresholds of statistical significance and therefore be submitted and published. Hence, an inverse relationship between sample size and effect size suggests the presence of bias in average effect sizes.ⁱⁱ Other studies that attempt to diagnose publication bias often use similar techniques. In virtually all cases, these studies focus on the relationship between the magnitude of effects and the size of the associated standard errors (e.g., Ashenfelter et al. 1999; Gorg and Strobl 2001; Stanley 2005).

In the present study we employ a similar approach. However, in contrast to many

previous examinations of publication bias, the literatures that we review examine a relatively broad class of effects using a variety of measures. Whereas a collection of studies of how a particular drug affects survival rates will all have similar treatments and outcome measures, researchers interested in economic voting may conceptualize and measure economic perceptions in different ways. They may also examine how these economic perceptions affect how citizens evaluate presidential candidates, candidates for the U.S. Congress, or those running for a seat in the state legislature. Furthermore, researchers may be interested in the degree to which economic voting is moderated by other factors, such as individuals' levels of political sophistication. Given the variety of effects addressed in the published work, we cannot simply look at the relationship between reported standard errors and effect sizes. Instead, we focus our analysis on a simple alternative test that measures how z-scores are distributed around the commonly accepted threshold of statistical significance.

The logic behind our approach is fairly intuitive. The sampling distribution that generates a reported coefficient estimate is assumed to be continuous. As such, if published results are unbiased, then we should expect to see roughly equal proportions of reported coefficients of interest just above and below *any* arbitrary value, and in particular, just above and below standard levels of statistical significance (i.e., p-values of .05). On the other hand, if these articles report an abundance of effects that barely exceed the standard threshold of statistical significance while reporting relatively few that fall just short of this threshold, this would be an anomaly and suggest publication bias.

Based on this logic, we employ a “caliper test” introduced by Gerber and Malhotra (2008a).ⁱⁱⁱ This test focuses on the distribution of reported z-scores for coefficients of interest around the accepted threshold of statistical significance. For example, for two-tailed tests we

examine z-scores that fall within +/- 10% of 1.96. We would expect that within this caliper z-scores should fall above and below the 1.96 threshold at approximately the same rate. If significantly more coefficients fall between 1.96 and 2.16 than fall between 1.76 and 1.96, then this implies the presence of publication bias.^{iv}

Section 2. Data

We identified the relevant articles published in ten political science journals: *American Journal of Political Science*, *American Political Science Review*, *American Politics Research*, *Journal of Politics*, *Political Behavior*, *Political Communication*, *Political Psychology*, *Political Research Quarterly*, *Public Opinion Quarterly*, and *Social Science Quarterly*. These journals were selected based on their prestige during the period we used as our sample frame and relevance to the current project. After identifying this pool of journals we used the Social Science Citation Index's key word search to locate articles relevant to each of our two areas of interest from 1990 to 2007.

Economic Voting Articles

The political science literature on economic voting analyzes how perceptions about the economy affect citizens' evaluations of political figures. A central issue addressed in this literature deals with the relative importance of pocketbook and sociotropic perceptions about the economy. In other words, which factors influence citizens' evaluations and voting decisions more: perceptions about their own personal economic situation or perceptions about the health of the economy as a whole? Over the years this literature also started to examine how other factors—such as political sophistication—might condition the relationships between these perceptions and vote choice.

Our search for articles on economic voting focused on three terms designed to identify any published research dealing with these questions. To ensure that we captured all relevant articles, we deliberately chose very broad search terms. The terms we searched for were: “economic voting,” “sociotropic,” and “pocketbook.” We captured all articles that included any of these terms in their abstracts, titles, or subject listings in the ten journals listed above. This search returned 57 articles, listed in the second column of Table A1. We double-checked our pool of articles by repeating this search in JSTOR for the years that were available (1990-2005).

The next step was to refine this list of 57 to a smaller list of topical articles that contained the needed information (coefficients and standard errors) to conduct the caliper test. We restricted our attention to articles about U.S. elections that analyzed voting or evaluations of candidates as the dependent variable and used sociotropic and/or pocketbook measures (and their moderators) as independent variables. Articles conducted using data on foreign countries,^v those analyzing how sociotropic/pocketbook perceptions affect evaluations other than those related to political candidates or voting, and those that did not publish standard errors were excluded from the analysis. This paring left us with 21 articles, listed in the third column of Table A1.

Last, we excluded articles that had a large number of hypotheses due to testing across several subgroups, years, regression specifications, and dependent variables. There are two rationales for this reduction. First, it minimizes the influence of any one article. Second, it is unclear what publication bias hypotheses predict for a paper with many coefficients. For example, Funk and Garcia-Monet (1997) present 80 coefficients and standards errors on economic variables across various models in their work. Including articles such as these would require judgment on our part as to which estimates were the most “important.” By restricting our analysis, we avoid the need to make such decisions.^{vi} We conducted our analysis using 19

articles, which are listed in the fourth column of Table A1. As discussed below, we assessed the sensitivity of our results to this culling process.

Negative Advertising Articles

Most of the literature on negative advertising examines whether negative advertising mobilizes turnout by evoking a sense that the election outcome matters or if it, instead, demobilizes potential voters from turning out by making them feel disenchanting with the political process. This literature is particularly interesting in the context of the present study. Some published work on the effects of negative advertising concludes that these ads depress turnout, while other work indicates that they stimulate turnout. It is not our goal to evaluate which of these findings is more valid. However, one consequence of mixed findings like these is that researchers do not have clear expectations about effect sizes and, as a result, are precluded from tailoring the power of their designs to most efficiently demonstrate statistically significant relationships. Thus, the negative advertising literature presents a rigorous test for the presence of publication bias in political science journals.

As with our search for articles on economic voting, our search for articles on negative advertising was designed to capture all relevant articles. We used seven broad search terms: “negative advertising,” “negative ads,” “negative advertisements,” “negative campaigning,” “negative campaign advertising,” “negative campaign ads,” and “negative campaign advertisements.” This search yielded 36 articles. In 2007, Lau et al. (2007) updated their earlier meta-analysis (Lau et al. 1999) of the negative advertising literature. We used their meta-analysis to ensure that our search captured the full range of published articles on the effects of negative advertising. Four articles were included in their meta-analysis that our search terms did not locate, leaving us with 40 articles in total, which are listed in the second column of Table A2.

We pared down this list of 40 articles to a list of topical articles that contained coefficients and standard errors. Articles dealing with candidate decisions to air negative ads (i.e., that used negative ads as a dependent variable), those that dealt with the views of children concerning negative ads (Rahn and Hirshorn 1999) and those that did not publish standard errors were excluded from the analysis.^{vii} This left us with 20 articles, listed in the third column of Table A2. Last, as with the articles on economic voting, we excluded articles with a large number of coefficients.^{viii} This truncation left us with 16 articles, listed in the fourth column of Table A2.

Selecting Coefficients

We recorded the z-statistics from all coefficients representing concepts of interest from each of these two pools of articles. For the economic voting literature we recorded the z-statistics for coefficients on all independent variables that measured either pocketbook or sociotropic attitudes, as well as coefficients that captured interactive (conditional) relationships. For the negative advertising literature we recorded z-statistics for coefficients on independent variables related to exposure to negative advertising and moderators of this effect. In both cases we recorded all coefficients across all regression specifications related to the topics of interest.

We illustrate our approach to coefficient selection using King and McConnell's (2003) study of the effects of negative advertising as an example. These authors conducted an experiment in the context of the 1996 Illinois Senate race in which treatment groups were exposed to varying numbers of negative campaign advertisements about the Republican candidate. The authors measured the impact of these treatments on affect towards each candidate as well as vote choice. The authors were also concerned with both the nonlinear effects of advertising and the moderating role of gender.

Table 1 of King and McConnell’s article presents regression results from the overall sample (2003, 852). There are three dependent variables: affect towards the Democratic candidate, affect towards the Republican candidate, and vote choice. The two variables dealing with the treatment of negative advertising are “Number of ads viewed” and “Number of ads viewed squared.” Hence, we recorded six coefficients from Table 1. Table 2 of the King and McConnell article includes interaction terms with gender to assess whether the effect of ads on women are different than their effect on men (853). Again, the authors estimate three regression specifications, one for each of the three dependent variables. The four variables dealing with the treatment of negative advertising are “Number of ads viewed,” “Number of ads viewed squared,” “Number of ads x gender,” and “Ads squared x gender.” Hence, we recorded an additional 12 coefficients from Table 2, making the total number of coefficients recorded from King and McConnell 18.

This process yielded 243 coefficients and standard errors related to economic voting and 149 related to the effects of negative advertising. For each z-score we also recorded whether the authors specified the relevant hypothesis test as one- or two-tailed.

Section 3. Results

We analyze the findings presented in each of the two literatures separately. Given that one- and two-tailed hypothesis tests imply different thresholds of significance, we present the distribution of z-scores from each of these two types of tests separately.

Figure 1 shows the distribution of the absolute values of z-scores for coefficients that specified a one-tailed test from the literature on economic voting. The width of the bars is set to 0.16 units—approximately 10% of 1.64 (1.64 corresponds to a p-value of .05). The figure shows a pronounced difference in the number of reported z-scores that fall just above and just below the

1.64 threshold. Twelve z-scores fall between 1.64 and 1.80; only 7 fall between 1.48 and 1.64. Figure 2 shows a similar pattern for the z-scores of the coefficients from studies of economic voting that evaluated hypotheses using a two-tailed test. Here the bars are 0.20 units—approximately 10% of 1.96—wide. Nine of these z-scores fall between 1.96 and 2.16. Only 3 fall between 1.76 and 1.96.^{ix}

[FIGURE 1 ABOUT HERE]

[FIGURE 2 ABOUT HERE]

Figures 3 and 4 show comparable z-score distributions from the literature on negative advertising. For one-tailed tests (displayed in Figure 3), only two z-scores fall just short of statistical significance, while five just barely reach significance. Similarly, in Figure 4 only four scores related to two-tailed tests fall between 1.76 and 1.96 while thirteen fall between 1.96 and 2.16.

[FIGURE 3 ABOUT HERE]

[FIGURE 4 ABOUT HERE]

One especially noteworthy aspect of the figures is that the number of cases in the interval just over the critical value is greater than the number falling in any other interval in three of the four cases (the exception being Figure 2), whereas the interval just below the critical value typically has very few z-scores. Overall, the ratio of economic voting results just over the critical values to those just under the critical values is about 2:1 for the 10% caliper. The results presented in the negative advertising literature are similar and the comparable ratio is 3:1.

We also performed similar analyses using wider (i.e., 15 and 20%) calipers. The results of this analysis, which pools one- and two-sided hypothesis tests, are presented in Table 1. For each caliper width we calculate the likelihood that the observed proportions of z-scores just

above and below the critical value are due to chance. For both literatures the data indicate that it is unlikely that the observed patterns are simply due to chance.^x

[TABLE 1 ABOUT HERE]

Next, we explored whether bias is most present in the most prominent political science journals. Although insignificant results may not be published in the discipline's most high-profile outlets, they may find their way to top subfield journals. Indeed, we do find evidence for this phenomenon. Examining only those articles published in the *American Political Science Review*, the *American Journal of Political Science*, and *The Journal of Politics*, across the two literatures we observe a 28:9 imbalance when applying the 10% caliper.^{xi} Conversely, for results published in all other journals, the ratio between coefficient estimates above and below the threshold is much more uniform (11:7). Although this suggests that many insignificant results eventually find their way into print, publication bias appears to be most prevalent in the most influential and cited journals.

We conducted a series of robustness checks to further test the findings. Perhaps our elimination of studies with a very large number of coefficients restricts the sample to studies with underspecified models. However, when we include studies with a large number of coefficient estimates (i.e., all articles that are topical and complete), the 10% caliper produces a ratio of 58:34 (pooling the two literatures), similar to the ratios reported in Table 1.^{xii} We also tried including the studies with large numbers of coefficients, but randomly selected coefficients such that the total would not exceed the cutoffs described above. Consistent with our previously reported findings, we obtained a 43:19 ratio (for the 10% caliper) when applying a cutoff of 32 coefficients for the economic voting literature and 25 coefficients for the negative advertising literature. On the other hand, the purest tests may be those papers that commit not only to a small

number of hypotheses, but also a limited number of coefficients associated with those hypotheses. Restricting the 10% caliper test to include only studies that test less than 10 coefficients, we again find a significant imbalance for the two literatures (9:2).

One additional issue to consider when evaluating these data is that while the critical value of .05 is widely accepted as the threshold of statistical significance, the z-score corresponding to this threshold is contingent on the researchers' judgment regarding whether the appropriate test is one- or two-tailed. For the significance levels to be valid, this determination must be independent from the estimation results. Selection of a one- or two-sided test is often a matter of discretion. It is possible that researchers, with the best of intentions and without any conscious effect on the results, may tend to conclude that it is more appropriate to report findings based on a one-tailed test when z-scores fall just short of the critical value for a two-tailed test.

We examine this possibility by applying the caliper test for two-tailed hypothesis testing (i.e., centered on the 1.96 threshold) to the sample of z-scores reported for one-tailed tests. Comparing the number of results in the interval between 1.64 and 1.96 with the equal sized interval over 1.96 allows us to examine whether there is a disproportionate number of cases evaluated by the one-tailed test that happen to pass this test but fail the more demanding two-tailed test. For results which fall in the lower portion of the caliper, it is especially advantageous to select a one-tailed test if it is anticipated that it is important to achieve statistical significance.

Table 2 reports the results of these tests. The findings suggest that the choice of hypothesis test may not be independent of the estimation results. Among the one-tailed tests reported in the economic voting literature 14 z-scores fall within the 1.64 and 1.96 under-caliper range while only three fall within the comparable over-caliper range. The p-value of .006 indicates that this ratio is highly unlikely to be due to chance. Although the corresponding test

for the negative advertising literature is not statistically significant, almost twice as many z-scores fall in the below-caliper range as fall in the above-caliper range. When we pool the two literatures we find that the number of z-scores falling in the under-caliper range significantly exceeds what we would expect to observe if these results were truly a representative sample of findings in these areas.^{xiii} These findings are consistent with the conclusion that decisions about what hypothesis tests to apply are not independent of the z-scores, which further suggests that the probability of a Type I error is likely to be underestimated and the reported p-values are incorrect.

[TABLE 2 ABOUT HERE]

Section 4. Discussion

The results suggest the presence of publication bias in two political behavior literatures, and that this bias is most prevalent in the leading journals. Of course, our tests may be underestimating the level of publication bias since in the later stages of a literature there may be greater incentive to uncover a statistically insignificant effect, potentially through several of the same practices described above (e.g., subgroup and model specification selection). Future research can explore this possibility in greater depth by examining literatures as they develop over time.

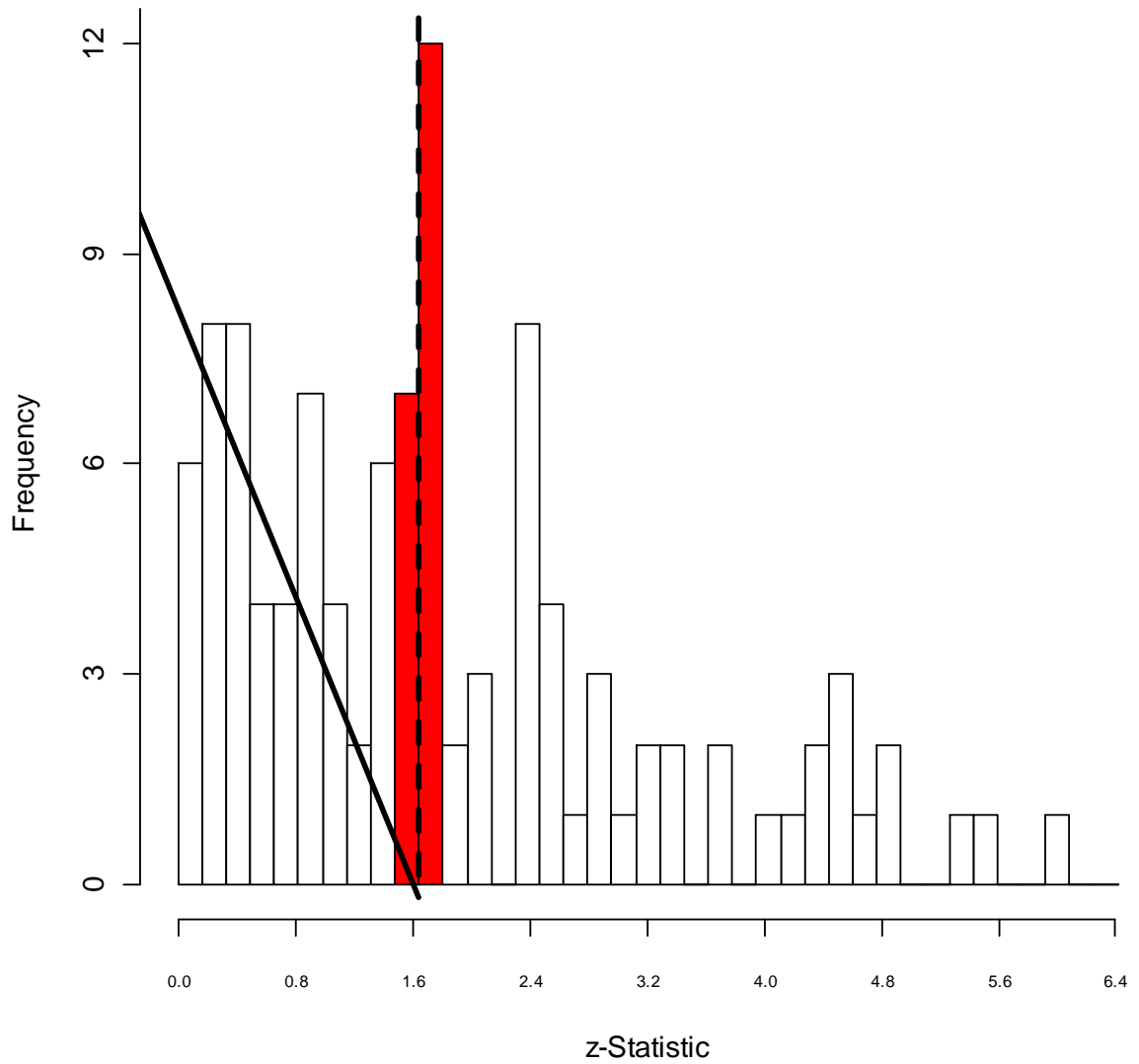
Gerber and Malhotra (2008a, 2008b) propose several potential responses to the problem of publication bias, including the establishment of study registries for political science. Our findings in this paper suggest that some further reflection on scholarly practices may be warranted. First, perhaps the chief constructive implication of our results is that political scientists in their roles as reviewers and authors should place more emphasis on research *design* and less emphasis on the $p < .05$ threshold. Undeniably, many insignificant results are the product

of poor research design (e.g., poor measurement, flawed administration of treatments, etc.). But consider the case of a study that produces a well-identified, but noisy estimate that fails to achieve statistical significance. Such work should be valued for what it contributes to the cumulative evidence on a question rather than dismissed because as a standalone study it is not sufficiently dispositive. Second, our finding that it appears that sometimes there is a switch from two-tailed to one-tailed hypothesis tests based on the obtained p-value suggests that scholars—at least so that they are clear in their own minds about the likelihood their findings are due to chance—should commit to a hypothesis test before collecting data and conducting analyses. Finally, authors can be encouraged to report sensitivity tests across multiple specifications, as is done in economics.

The influence of journal practices on scholarship is a common topic of discussion when researchers gather in informal settings, but is rarely a subject for empirical inquiry. There are strong opinions about statistical reporting conventions that have prompted experiments with alternative publication practices. Some scholars have questioned the value of hypothesis testing and discourage reliance upon hypothesis tests and p-values (Gill 1999; Fidler et al. 2004). In one extreme case, an editor banned p-values from the journal during his editorship (Rothman 1998). The advisability of this or other measures is ultimately an empirical question. However, there is unfortunately very little available research on the incentive effects of alternative journal standards and practices in political science. While we focus on a specific type of publication bias in this article, our work is a small contribution to this broader effort of understanding how the production of scholarly research is shaped by incentives. Although our goal here was not to quantify the degree to which these two literatures are biased, meta-analyses of particular literatures can reveal how sensitive well-understood findings are to the possibility of publication

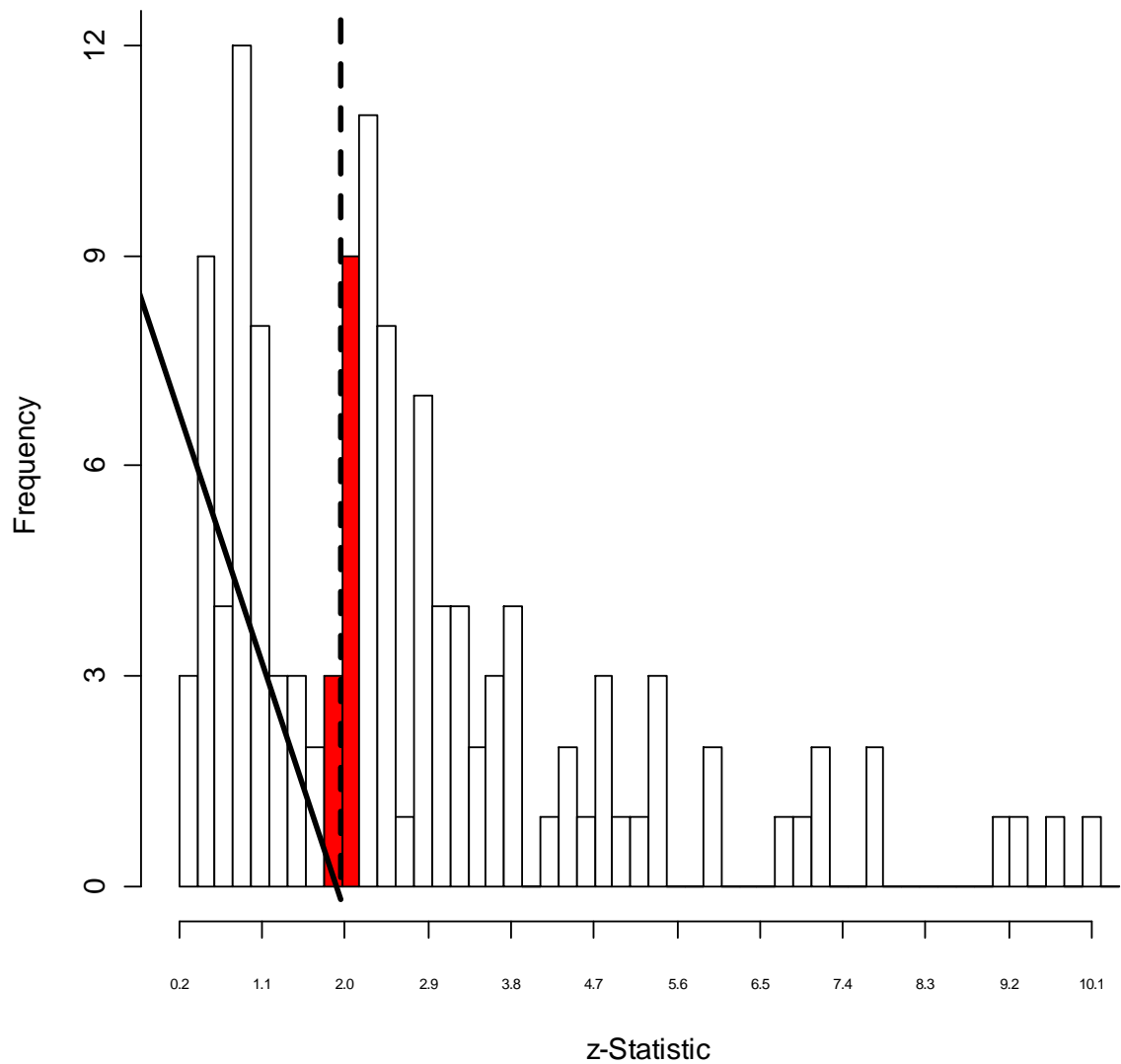
bias (e.g., Pettigrew and Tropp 2006). Our evidence suggesting research is affected by reporting conventions indicates that understanding how scholarship is affected by the incentive environment more generally is a fruitful topic for further research.

Figure 1: Histogram of z-Statistics, Economic Voting (One-Tailed)



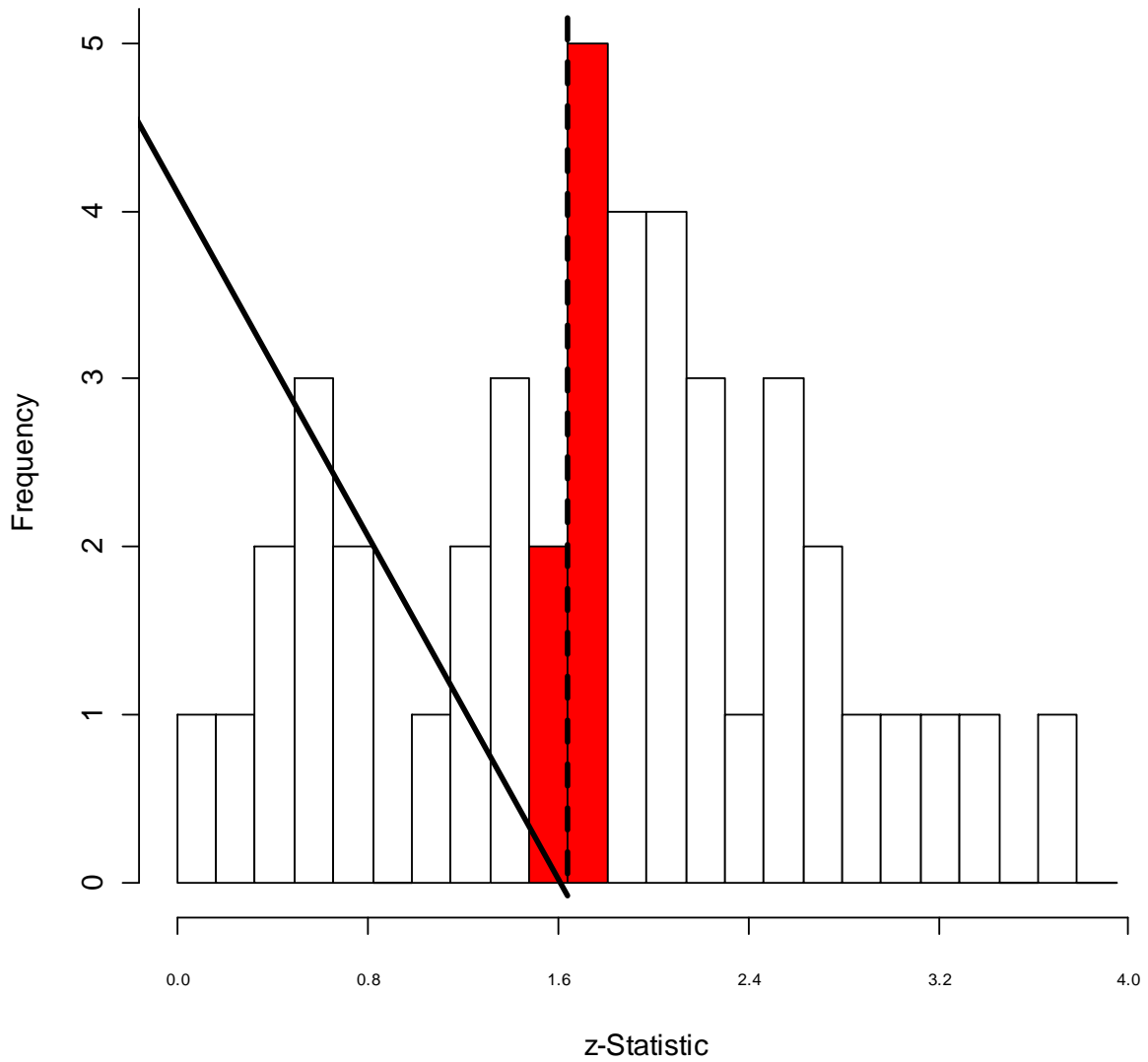
Note: Width of bars (.16) approximately represents 10% caliper. Dotted line represents critical z-statistic (1.64) associated with $p=.05$ significance level for one-tailed tests.

Figure 2: Histogram of z-Statistics, Economic Voting (Two-Tailed)



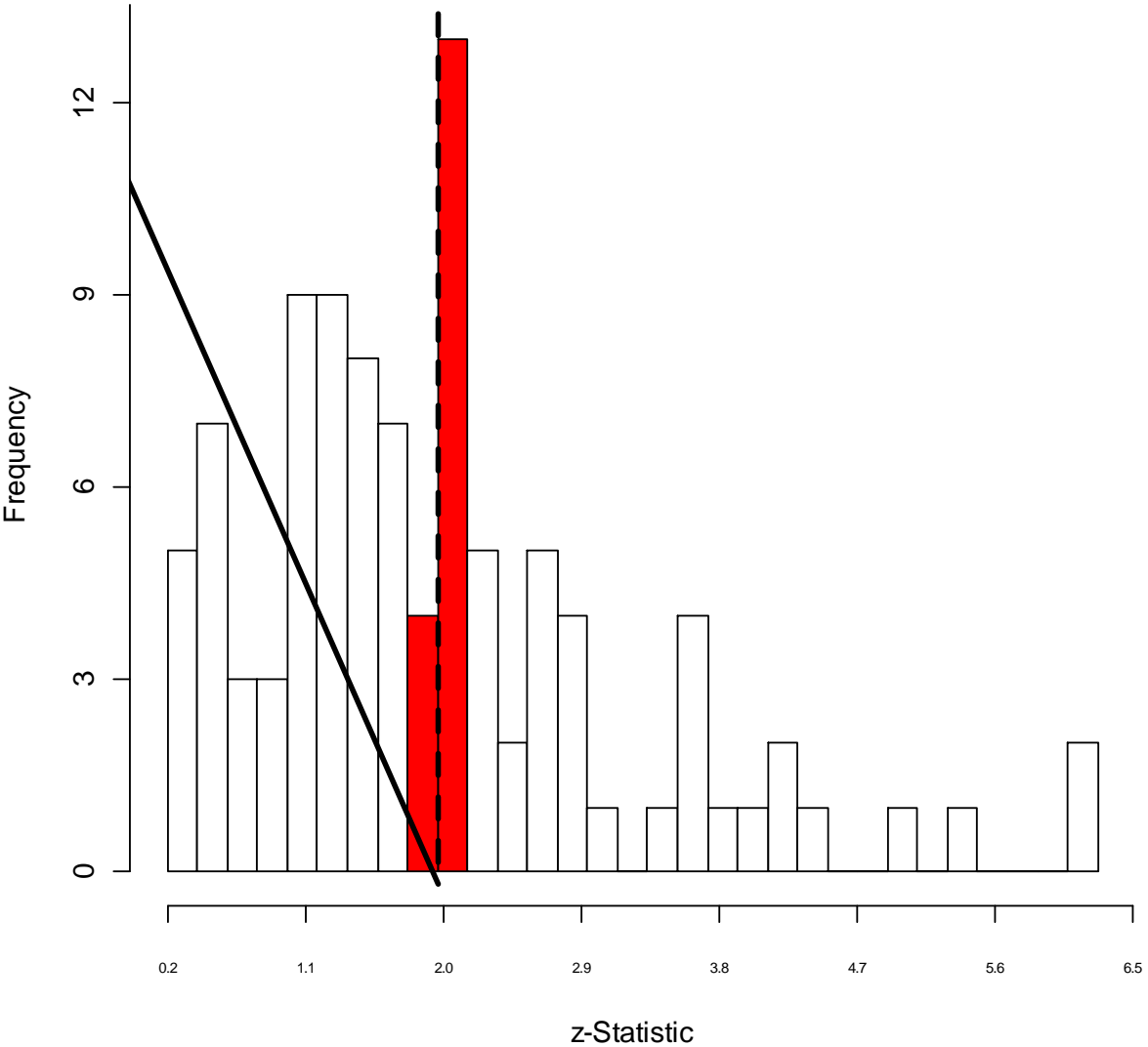
Note: Width of bars (.2) approximately represents 10% caliper. Dotted line represents critical z-statistic (1.96) associated with $p=.05$ significance level for two-tailed tests.

Figure 3: Histogram of z-Statistics, Negative Advertisements (One-Tailed)



Note: Width of bars (.16) approximately represents 10% caliper. Dotted line represents critical z-statistic (1.64) associated with $p=.05$ significance level for one-tailed tests.

Figure 4: Histogram of z-Statistics, Negative Advertisements (Two-Tailed)



Note: Width of bars (.2) approximately represents 10% caliper. Dotted line represents critical z-statistic (1.96) associated with $p=.05$ significance level for two-tailed tests.

Table 1: Caliper Tests of Publication Bias in Economic Voting and Negative Advertising Literatures

	<u>Over Caliper</u>	<u>Under Caliper</u>	<u>p-value</u> *
<u>Economic Voting Literature</u>			
10% Caliper	21	10	.035
15% Caliper	28	16	.048
20% Caliper	34	18	.018
<u>Negative Ads Literature</u>			
10% Caliper	18	6	.011
15% Caliper	25	11	.014
20% Caliper	27	16	.063

*Based on density of binomial distribution (two-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 for one-tailed tests and 1.76-1.96 for two-tailed tests.

Table 2: Caliper Tests of Publication Bias in Economic Voting and Negative Advertising Literatures – reported 1-tailed tests, 2-tailed caliper

	<u>Over Caliper</u>	<u>Under Caliper</u>	<u>p-value</u> *
<u>Economic Voting Literature</u>			
15% Caliper	3	14	.006
<u>Negative Ads Literature</u>			
15% Caliper	5	9	.212
<u>Pooled</u>			
15% Caliper	8	23	.005

*Based on density of binomial distribution (two-tailed)

Note: “Under Caliper” indicates number of results with z-scores between 1.64 and 1.96; “Over Caliper” indicates the number of results between 1.96 and 2.28.

References

- Ashenfelter, O., Harmon C., and Oosterbeek, H. (1999). "A Review of Estimates of the Schooling/ Earnings Relationship, with Tests for Publication Bias." *Labour Economics*, 6, 453-470.
- Card, D., and Krueger, A. B. (1995). "Time-Series Minimum Wage Studies: A Meta-Analysis." *American Economic Review*, 85, 238-243.
- Coursol, A., and Wagner, E. E. (1986). "Effect of Positive Findings on Submission and Acceptance Rates: A Note on Meta-Analysis Bias." *Professional Psychology*, 17, 136-137.
- De Long, J. B., and Lang, K. (1992). "Are All Economic Hypotheses False?" *Journal of Political Economy*, 100, 1257-1272.
- Doucouliafos, C. (2005). "Publication Bias in the Economic Freedom and Economic Growth Literature." *Journal of Economic Surveys*, 19, 367-387.
- Doucouliafos, C., Laroche, P., and Stanley, T. (2005). "Publication Bias in Union-Productivity Research?" *Industrial Relations*, 60, 320-347.
- Fidler, F., Cumming, G., Burgman, M., and Thomason, N. (2004). "Statistical Reform in Medicine, Psychology and Ecology." *Journal of Socio-Economics*, 33, 615-630.
- Funk, C. L., and Garcia-Monet, P. A. (1997). "The Relationship between Personal and Nation Concerns in Public Perceptions about the Economy." *Political Research Quarterly*, 50, 317-342.
- Gerber, A. S., Green, D. P., and Nickerson, D. (2000). "Testing for Publication Bias in Political Science." *Political Analysis*, 9, 385-392.
- Gerber, A. S., and Malhotra, N. (2008a). "Publication Incentives and Empirical Research: Do

- Reporting Standards Distort the Published Results?" *Sociological Methods and Research*, 37, 3-30.
- Gerber, A. S., and Malhotra, N. (2008b). "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science*, 3, 313-326.
- Gill, J. (1999). "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly*, 52, 647-674.
- Gorg, H., and Strobl, E. (2001). "Multinational Companies and Productivity: A Meta-Analysis." *The Economic Journal*, 111, 723-739.
- Gotzsche, P. C. (2006). "Believability of Relative Risks and Odds Ratios in Abstracts: Cross Sectional Study." *British Medical Journal*, 333, 231-234.
- Greenwald, A. G. (1975). "Consequences of Prejudice Against the Null Hypothesis." *Psychological Bulletin*, 82, 1-20.
- King, J. D., and McConnell, J. B. (2003). "The Effect of Negative Campaign Advertising on Vote Choice: The Mediating Influence of Gender." *Social Science Quarterly*, 84, 843-857.
- Lau, R. R., Sigelman, L., Heldman, C., and Babbitt, P. (1999). "The Effects of Negative Political Advertisements: A Meta-Analytic Assessment." *American Political Science Review*, 93, 851-875.
- Lau, R. R., Sigelman, L., and Rovner, I. B. (2007). "The Effects of Negative Political Campaigns: A Meta-Analytic Reassessment." *Journal of Politics*, 69, 1176-1209.
- Pettigrew, T. F., and Tropp, L. R. (2006). "A Meta-Analytic Test of Intergroup Contact Theory." *Journal of Personality and Social Psychology*, 90, 751-783.

- Plümper, T. (2007). "Academic Heavy-Weights: The 'Relevance' of Political Science Journals." *European Political Science*, 6, 41-50.
- Rahn, W. M., and Hirshorn, R. M. (1999). "Political Advertising and Public Mood: A Study of Children's Political Orientations." *Political Communication*, 16, 387-407.
- Rothman, K. J. (1998). "Writing for Epidemiology." *Epidemiology*, 9, 333-337.
- Stanley, T. D. (2005). "Beyond Publication Bias." *Journal of Economic Surveys*, 19, 309-337.
- Sterling, T. D. (1959). "Publication Decision and the Possible Effects on Inferences Drawn From Tests of Significance—or Vice Versa." *Journal of the American Statistical Association*, 54, 30-34.

Appendix

Table A1: Economic Voting Literature			
Journal	Search Results	Topical and Complete	Test < 33 Coefficients
<i>American Journal of Political Science</i>	Gomez and Wilson (2001) Quinn and Woolley (2001) Duch et al. (2000) Alvarez and Nagler (1998) Krause (1997) Mutz and Mondak (1997) Hetherington (1996) Pacek and Radcliff (1995) Alvarez and Nagler (1995) Clarke and Stewart (1994) Pacek (1994) Powell and Whitten (1993) Mutz (1992) Suzuki (1991)	Gomez and Wilson (2001) - 22 Alvarez and Nagler (1998) - 4 Mutz and Mondak (1997) - 14 Hetherington (1996) - 6 Alvarez and Nagler (1995) - 4 Mutz (1992) - 25	Gomez and Wilson (2001) Alvarez and Nagler (1998) Mutz and Mondak (1997) Hetherington (1996) Alvarez and Nagler (1995) Mutz (1992)
<i>American Political Science Review</i>	Basinger and Lavine (2005) Duch (2001) Roberts and Wibbels (1999) Kaufman and Zuckermann (1998) Radcliff (1992)	Basinger and Lavine (2005) - 16	Basinger and Lavine (2005)
<i>American Politics Research</i>	Barker and Muraca (2003) Lockerbie (2002) Rudalevige (2001)		
<i>Journal of Politics</i>	Arce (2003) Hellwig (2001) Norpoth (2001) Nadeau and Lewis-Beck (2001) Gartner and Segura (2000) Shah et al. (1999) Kahn and Kenney (1997) Suzuki and Chappell (1996) Fackler and Lin (1995) Pacek and Radcliff (1995) Welch and Hibbing (1992) Sigelman (1991) Stein (1990)	Nadeau and Lewis-Beck (2001) - 3 Welch and Hibbing (1992) - 16 Stein (1990) - 12	Nadeau and Lewis-Beck (2001) Welch and Hibbing (1992) Stein (1990)
<i>Political Behavior</i>	Weisberg (2002) Weatherford and Sergeev (2000) Niemi et al. (1999) Books and Prysby (1999) Wlezien et al. (1997) Lanoue (1991) Sigelman et al. (1991) Lau et. al (1990)	Weatherford and Sergeev (2000) - 16 Books and Prysby (1999) - 11 Lanoue (1991) - 6	Weatherford and Sergeev (2000) Books and Prysby (1999) Lanoue (1991)
<i>Political Communication</i>	No articles	No articles	No articles
<i>Political Psychology</i>	No articles	No articles	No articles
<i>Political Research</i>	Gomez and Wilson (2003) Arcenaux (2003)	Godbout and Belanger (2007) - 104 Glasgow (2005) - 6	Glasgow (2005) Gomez and Wilson (2003)

<i>Quarterly</i>	Rudolph and Grant (2002) Bohrer and Tan (2000) Weyland (1998) Chaney et al. (1998) Goren (1997) Funk and Garcia-Monet (1997) Romero and Stambough (1996) Holbrook and Garand (1996) Radcliff (1994) Lanoue (1994)	Gomez and Wilson (2003) - 32 Rudolph and Grant (2002) - 5 Chaney et al. (1998) - 18 Goren (1997) - 24 Funk and Garcia-Monet (1997) - 80 Romero and Stambough (1996) - 3	Rudolph and Grant (2002) Chaney et al. (1998) Goren (1997) Romero and Stambough (1996)
<i>Public Opinion Quarterly</i>	No articles	No articles	No articles
<i>Social Science Quarterly</i>	Joslyn and Haider-Markel (2007) Caplan (2002)	No articles	No articles
Note: The "Topical and Complete" column reports the number of coefficients recorded from each article.			

Table A2: Negative Advertising Literature			
Journal	Search Results	Topical and Complete	Test < 26 Coefficients
<i>American Journal of Political Science</i>	Brooks and Geer (2007)* Brader (2005)* Lau and Pomper (2002) Freedman and Goldstein (1999) Finkel and Geer (1998) Brians and Wattenberg (1996)	Lau and Pomper (2002) - 54 Freedman and Goldstein (1999) - 4 Finkel and Geer (1998) - 86 Brians and Wattenberg (1996) - 6	Freedman and Goldstein (1999) Brians and Wattenberg (1996)
<i>American Political Science Review</i>	Ansolabehere et al. (1999) Kahn and Kenney (1999) Wattenberg and Brians (1999) Skaperdas and Grofman (1995) Ansolabehere et al. (1994)	Ansolabehere et al. (1999) - 14 Kahn and Kenney (1999) - 7 Wattenberg and Brians (1999) - 2 Ansolabehere et al. (1994) - 6	Ansolabehere et al. (1999) Kahn and Kenney (1999) Wattenberg and Brians (1999) Ansolabehere et al. (1994)
<i>American Politics Research</i>	Krebs and Holian (2007) Herrnsom and Lucas (2006) Fridkin and Kenney (2004)	Fridkin and Kenney (2004) - 48	
<i>Journal of Politics</i>	Brooks (2006) Clinton and Lapinski (2004) Sigelman and Buell (2003) Sigelman and Kugler (2003) Goldstein and Freedman (2002) Sigelman and Shiraev (2002) Lau and Pomper (2001) Theilmann and Wilhite (1998)	Brooks (2006) - 6 Clinton and Lapinski (2004) - 25 Goldstein and Freedman (2002) - 1 Lau and Pomper (2001) - 13	Brooks (2006) Clinton and Lapinski (2004) Goldstein and Freedman (2002) Lau and Pomper (2001)
<i>Political Behavior</i>	Kahn and Geer (1994)	No articles	No articles
<i>Political Communication</i>	Craig, Kane, and Gainous (2005) Leshner and Thorson (2000) Rahn and Hirshorn (1999) Klotz (1998) Hitchon et al. (1997)* Kern and Just (1995) Tinkham and Weaver Lariscy (1995)	Craig, Kane, and Gainous (2005) - 8 Hitchon et al. (1997)* - 10	Craig, Kane, and Gainous (2005) Hitchon et al. (1997)*
<i>Political Psychology</i>	Martin (2004) Schultz and Pancer (1997)*	Martin (2004) - 15	Martin (2004)
<i>Political Research Quarterly</i>	Niven (2006) Sanders and Norris (2005) Stevens (2005) Djupe and Peterson (2002) Damore (2002)	Sanders and Norris (2005) - 12 Stevens (2005) - 42 Djupe and Peterson (2002) - 2	Sanders and Norris (2005) Djupe and Peterson (2002)
<i>Public Opinion Quarterly</i>	No articles	No articles	No articles
<i>Social Science Quarterly</i>	King and McConnell (2003) Hale et al. (1996) Kaid et al. (1993)	King and McConnell (2003) - 18	King and McConnell (2003)
<i>Note:</i> The "Topical and Complete" column reports the number of coefficients recorded from each article. * From Lau et al. (2007)			

ⁱ Since these steps are not explicitly part of post-submission publication decisions, they contribute to what may more precisely be called “specification bias.”

ⁱⁱ An alternative interpretation, which we and the authors view as implausible, is that the true effects varied considerably over time and across modes of communication in a fashion that happened to match the large variations in sample sizes.

ⁱⁱⁱ A more detailed discussion and formal presentation of this test is presented in Gerber and Malhotra (2008a).

^{iv} Following Gerber and Malhotra (2008a), we assume that the asymptotic sampling distribution of z , $F(z)$, is continuous. This suggests that no matter what the true effect is (whether it is small or large), over any narrow region the conditional probability of observing an outcome that falls in a subset in an interval is approximately equal to the relative proportion of the subset to the interval.

^v Studies of foreign countries were considered part of a separate literature since economic voting may vary with cultural, political, and institutional context. For example, blame and credit attribution may depend on a variety of country-specific factors such as the electoral system, the ability of the government to control the domestic economy, and the reliance of the country on foreign trade. An analysis of publication bias in the literature examining economic voting outside of the United States would be a fruitful avenue for future research.

^{vi} The maximum number of coefficients we allowed was 32, a threshold which excluded 2 articles. This value formed a natural discontinuity as the next article tested 80 coefficients. Thirty-two could be considered a large number of hypotheses tested in one article as well. We have chosen to err on the side of caution (i.e., bias ourselves *against* finding evidence of publication bias) and only exclude articles that are clearly outliers in terms of their number of

coefficients. Thus, our results should understate the amount of publication bias. Using different cutoffs yielded results similar to the ones presented below for both the economic voting and negative advertising literatures.

^{vii} Three of the four articles identified by Lau et al. (2007) that were not captured by our search were excluded from our analysis either because they did not present standard errors or because they address effects outside of our identified area of interest.

^{viii} The maximum number of coefficients we allowed was 25, which, again, formed a natural discontinuity.

^{ix} The results are qualitatively similar when we vary the size of the intervals on either side of the critical values. See Table 1 for details.

^x One factor that complicates the analysis presented in Table 1 is that some studies contribute more than one coefficient in the caliper, suggesting that each coefficient cannot be viewed as statistically independent. Although the departure from independence over the narrow range of values included in the caliper is almost certainly trivial, we performed robustness checks by restricting attention to those studies that contribute only one or two coefficients (i.e., where there should not be an issue of nonindependent observations). We also observe an imbalance among these studies. Pooling the two literatures, we find 14 studies contribute one (10 studies) or two (4 studies) coefficients for the 10% caliper with 15 coefficients just over the critical value and only 3 just under. The likelihood of such an imbalance (15:3), under the hypothesis of equal probability, is less than 0.004.

^{xi} *APSR*, *AJPS*, and *JOP* rank first, second, and fifth, respectively, according to Thomson Scientific's *Journal Performance Indicators* (<http://in-cites.com/rsg/jpi/>) for the period 1981-2007. *Public Opinion Quarterly* ranks third due to citation in fields outside of political science

(e.g. psychology, sociology, survey methods). However, *POQ* contributes no coefficients to our analyses. The next highest ranked journal from our list is *American Politics Research* (formerly *American Politics Quarterly*) at number 10. To see the top ten political science journals based on this ranking system, visit http://sciencewatch.com/dr/sci/09/mar29-09_1/. See Plümper (2007, Table 4) for a comparison of various journal ranking systems.

^{xii} The ratio for each literature is 27:18 for economic voting and 31:16 for negative advertising.

^{xiii} The ratios are quite similar if we include coefficients from all the articles: 6:18 for economic voting and 8:14 for negative advertising.