Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?

Alan S. Gerber and Neil Malhotra

Sociological Methods Research 2008; 37; 3 originally published online Jun 10, 2008;
DOI: 10.1177/0049124108318973

Sociological Methods & Research
Volume 37 Number 1
August 2008 3-30

10.1177/0049124108318973
http://smr.sagepub.com
hosted at
http://online.sagepub.com

Publication Bias in Empirical Sociological Research

Do Arbitrary Significance Levels Distort Published Results?

Alan S. Gerber
Yale University, New Haven, Connecticut
Neil Malhotra
Stanford University, Stanford, California

Despite great attention to the quality of research methods in individual studies, if publication decisions of journals are a function of the statistical significance of research findings, the published literature as a whole may not produce accurate measures of true effects. This article examines the two most prominent sociology journals (the *American Sociological Review* and the *American Journal of Sociology*) and another important though less influential journal (*The Sociological Quarterly*) for evidence of publication bias. The effect of the .05 significance level on the pattern of published findings is examined using a "caliper" test, and the hypothesis of no publication bias can be rejected at approximately the 1 in 10 million level. Findings suggest that some of the results reported in leading sociology journals may be misleading and inaccurate due to publication bias. Some reasons for publication bias and proposed reforms to reduce its impact on research are also discussed.

Keywords: *publication bias; caliper test; meta-analysis; hypothesis testing*

Most empirical research in sociology aims to produce accurate measurement of causal effects. Throughout the social sciences there is

Authors' Note: We acknowledge Andrew Gelman, Donald Green, Alexander Tahk, Jowei Chen, Alexander Kuo, Sarah Anzia, Jeremy Freese, Sean Riordan, and Kendra Bischoff for valuable suggestions. We also thank Ray Selie and Tania Juarez for helpful research assistance. Please address correspondence to Neil Malhotra, Stanford University, Department of Political Science, Encina Hall West, Room 100, Stanford, CA 94305-6044; e-mail: neilm@stanford.edu. An online appendix is available at http://smr.sagepub.com/supplemental.

an increasing appreciation for how difficult it is to separate causal relationships from spurious ones. Simple regression methods are being supplemented by methodological advances in observational research such as the use of panel data and matching methods (e.g., Correll 2001; DiPrete and Engelhardt 2004; Halaby 2004). Perhaps more important, there is greater emphasis on natural experiments and regression discontinuity designs along with increased use of experimental methods in laboratory and especially in naturalistic contexts (e.g., Pager 2003; for a demonstration of this development within a literature, see the review by Mouw 2006). Together, these advances have made it much more plausible, compared to earlier decades, to interpret the results of studies as unbiased estimates of causal effects rather than correlations.

Unfortunately, improving the research methods used in individual studies is not sufficient to ensure that the collections of studies that form sociological literatures yield valid conclusions. If some findings are more likely to be published than others, literatures will be biased even if each study appears methodologically rigorous and convincing. Furthermore, publication bias may induce biased research by exerting subtle pressures that affect the implementation of stated research protocols, which, if they had been executed in an evenhanded manner, would lead to unbiased results. Literatures are quite vulnerable to publication bias since although studies are reviewed one at a time, bias becomes evident only when the collection of published results is examined. It is necessary to step back and consider the pattern of published results to see the cumulative consequences of individual accept and reject decisions.

This article undertakes this task and examines the publication record of leading sociology journals to see if there is evidence of publication bias. Publication bias occurs when the probability that a result is published depends on the estimates produced by the study, holding the methodological quality of the study fixed. If a study's publication probability is affected by the study's estimates, then the published estimates will not be a draw from the true sampling distribution of the estimator being reported in the research, but rather some unknown function of it. The effects of publication bias depend on the form the bias takes, the attributes of the studies that might potentially be published, and the true value of the parameter being estimated. In one recent study of publication bias in the political science literature on voter turnout, Gerber, Green, and Nickerson (2000) show that the published literature is filled with reports of large, upwardly biased effects in research areas where studies tend to have small samples, the true effect being studied is small, and publication bias takes the form of rejecting statistically

insignificant findings. Interestingly, under these three conditions the genuine or underlying population parameter value has almost *no* effect at all on the results reported in the literature; the published "findings" are produced nearly entirely by sampling variability.

We adopt a broad interpretation of publication bias, which we define as the outcome that publication practices lead to bias in parameter estimates. This can be produced in several ways. First, editors and reviewers may prefer significant results and reject methodologically sound articles that do not achieve certain thresholds. Alternatively, scholars may send only studies with significant results to journals and place the rest in "file drawers," even if they are of high methodological quality. It is also possible that arbitrary significance levels encourage researchers to examine different model specifications and population subgroups to push results below certain thresholds. 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. This effect may be particularly pronounced in the social sciences, where the focus is on changes in point estimates rather than shifts in confidence intervals. 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 article evaluates the primary concern that the published results are not an accurate reflection of the research undertaken in the discipline.

Standard methods for detecting publication bias assess the pattern of results reported across studies measuring the effects of a given treatment. One sign that publication bias based on the .05 significance level is present in a literature is if smaller studies tend to report larger results. Without a large estimated treatment effect, which may be produced by chance, the ratio of estimated treatment effect to the standard error will not exceed the critical values needed to generate a statistically significant t ratio, and the paper will not be published or submitted. Alternatively, the researcher may collect more data until the significance threshold is broken. Gerber et al. (2000) examined the half dozen published field experiments measuring the effect of various modes of political communication on voter turnout and found evidence of this form of publication bias. However, a different method must be used to detect publication bias across studies of different effects 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 U.S. states).

In this article we conduct a broad examination of publication bias by considering all statistical studies on all topics published in the *American Sociological Review* (*ASR*), the *American Journal of Sociology* (*AJS*), and *The Sociological Quarterly* (*TSQ*) over the past three years.[1] We also examine three volumes from the *ASR* from one decade ago to see if the situation has changed over time. These three publications characterize a set of journals widely read by sociologists. Due to their influence in the field, readers place great weight on the findings in these journals, meaning that a detailed analysis of them is warranted. We use a simple, intuitive test for detecting the presence of publication bias recently introduced in political science by Gerber and Malhotra (2006). We examine the ratio of reported results just above and just below the critical value associated with the .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. Borrowing the logic of regression discontinuity design (Campbell 1969), there is no reason to expect that in the narrow region just above and below the critical value, there will be substantially more cases above than below the critical value unless the .05 level was somehow affecting what is being published. Our examination of the articles in the leading journals in sociology shows that across a range of definitions of "just above" and "just below" the critical value, 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 10 million. Our findings suggest that some of the results reported in the leading sociology journals may be misleading and inaccurate due to publication bias. After documenting this for the *ASR* and the *AJS* as well as *TSQ*, we speculate about the causes and consequences.

Understanding the extent and implications of publication bias in sociology is a vast undertaking, and that expansive subject is only part of the even larger question of how the research environment affects scholarly production. The aim of this article is to take an important step by providing convincing evidence of publication bias in the leading journals. Our discussion of the sources and implications of these findings, and the suggestion of several correctives, is much more speculative and tentative.

We want to stress up front that this article is not intended to be an indictment of the field of sociology, which has produced many theoretical and empirical contributions that have influenced researchers across the social sciences. We focus on sociology here not because publication bias

is an especially significant problem in the field as compared to other disciplines. Indeed, as discussed below, publication bias is ubiquitous across the natural and social sciences. Rather, we simply intend to raise this issue among the sociological community in hopes of further improving research quality in an already robust discipline.

The article is organized as follows. The next section reviews the literature on publication bias. The following section describes the statistical test we use to detect publication bias, the caliper test (Gerber and Malhotra 2006). Then we describe how we constructed the data set for our statistical analysis and present the results for the *ASR*, the *AJS*, and *TSQ*. Finally, we discuss possible objections to our analysis and the implications of our findings. We also suggest some methods for reducing publication bias and present ideas for further research.

Literature Review

In this section, we review the existing literature on publication bias. We begin with a discussion of techniques for detecting bias and then discuss studies that have found evidence of it in other social science disciplines. The fact that publication bias seems to be widespread across disparate fields suggests that an examination of sociology is warranted.

Methodological Approaches

Most of the statistical research on publication bias is in the context of meta-analysis, a widely used technique for summarizing evidence, usually experimental, from multiple studies in a particular literature (Sutton et al. 2000). The goal of meta-analysis is to summarize the overall treatment effect across a group of studies, taking advantage of the fact that pooling data increases the precision of the estimates. However, the results of meta-analyses of published studies may be biased since only the studies that are submitted to journals and survive the review process are typically included. If unpublished studies that show weak or insignificant treatment effects are excluded, then the overall treatment effect may be biased upward.

The most widely used method of detecting publication bias is a visual inspection of a funnel graph, which plots effect size against some measure of precision (e.g., the sample size or the inverse of the standard error; Light and Pillemar 1984). In the absence of publication bias, the graph

should look like a funnel; studies with large sample sizes should have treatment effects clustered near the average, whereas studies with small sample sizes should have widely dispersed effect sizes. Statistical tests of the funnel graph have been developed to objectively assess the degree of skewness in the shape of the funnel. Stanley and Jarrell (1989) and Egger et al. (1997) suggest a simple model regressing effect size against precision, using weighted least squares to correct for heteroskedasticity.[2] In the absence of publication bias, effect size should vary randomly around the true effect size (the intercept), independent of the degree of precision.[3]

The funnel graph and related methods can be used to detect publication bias in collections of studies that estimate a particular treatment effect. To detect publication bias associated with critical values for a general collection of studies, Gerber and Malhotra (2006) propose using what they call a "caliper test." They note that if critical values are affecting what is submitted or accepted for publication, the proportion of published results just over critical values will exceed that just under critical values to an extent that cannot be explained by chance. As our study uses this test, it will be discussed in greater detail in the next section of the article.

In addition to these statistical techniques for detecting publication bias, several qualitative methods have also been proposed. For instance, one can calculate the percentage of published studies that reject the null hypothesis and compare it to standard significance levels (e.g., 5%; Sterling 1959). The statistical significance of published studies can be compared to that of unpublished studies such as dissertations or conference presentations (Glass, McGaw, and Smith 1981). Finally, surveys of researchers, reviewers, and editors can be used to assess whether there is a bias against statistically insignificant results (Coursol and Wagner 1986).

Evidence of Publication Bias in Academic Disciplines

Investigations of publication bias are most prevalent in experimental clinical trials research, in which meta-analyses are common (Begg and Berlin 1988). For instance, Easterbrook et al. (1991) conduct an audit of studies approved by a clinical trials human subjects board over a four-year time period and find that studies with significant results were more likely to be published, resulted in more publications, and appeared in more prestigious journals. Berlin, Begg, and Louis (1989) apply the funnel graph methodology to a sample of clinical cancer trials and find a strong negative relationship between the sample size of the study and the effect size of the drug, controlling for covariates such as the use of randomization and the number of

research centers involved. For example, studies with 1 to 50 participants exhibit overall survival times nearly four times as large as studies with more than 101 participants.

Many other areas of medical research have been found to suffer from publication bias, which has been demonstrated to have had a material effect on important literatures. Simes (1986) compares published ovarian cancer trials to unpublished (yet registered) trials. The average p value of published studies is .02, whereas the average p value of unpublished studies is .24. Furthermore, the unpublished studies also have somewhat smaller effect sizes, meaning that the overall efficacy suggested by the literature is overstated. Using a funnel graph, Levois and Layard (1995) observe that the results of small sample studies are skewed toward finding that environmental tobacco smoke is harmful. However, a set of unpublished, registered studies shows no significant impact of secondhand smoke on coronary heart disease. Surveying obesity treatment studies, Allison, Faith, and Gorman (1996) regress effect sizes against standard errors and find a significant positive relationship, suggesting that the published literature overstates the efficacy of obesity treatment programs.

In the field of social work, experiments by Epstein (2000, 2004), in which a positive and a negative version of the same study were randomly sent to journals, showed a bias toward the positive version. Within the social sciences, Sterling (1959) was the first to conduct an audit of the four main psychology journals and found that in a single year, 97.3% of articles rejected the null hypothesis at the 5% significance level. In a recent audit of the same journals, Sterling et al. (1995) note that the situation had changed little since the 1950s, with 95.6% of articles rejecting the null. Surveys of psychologists have found that authors are reluctant to submit insignificant findings and journal editors are less likely to publish them (Greenwald 1975). Coursol and Wagner (1986) find that psychologists reported submitting 82.2% of studies that had significant findings but only 43.1% of studies with neutral or negative findings. In addition to this bias in the submission stage, researchers reported that 65.9% of significant studies submitted for publication were ultimately accepted, whereas only 21.5% of insignificant studies eventually appeared in print. Using a quantitative methodology that accounts for study characteristics, McLeod and Weisz (2004) compare unpublished dissertations on youth psychotherapy to published studies and observe that although the theses were methodologically superior to the published work, their effect sizes were half as large, even when controlling for methodological differences such as length of treatment sessions and therapist type.

In economics, publication bias is raised as a concern by De Long and Lang (1992), who conduct an audit of the four main economics journals and observe that *not a single* article found the null hypothesis to be "true" (i.e., all p values were below .90). Card and Krueger (1995) review the minimum wage effects literature and find a significant negative relationship between t statistics and sample sizes. Examining the returns to education literature, Ashenfelter, Harmon, and Oosterbeek (1999) find a strong positive relationship between effect size and standard errors—evidence of publication bias. Moreover, using the Hedges (1992) weight function, they find that results from published studies tend to fall below given threshold levels (e.g., $p = .01$ or $p = .05$). They then estimate a much lower true education effect than suggested by the literature.

Using Stanley and Jarrell's (1989) metaregression technique, Gorg and Strobl (2001) detect a positive relationship between coefficient size and standard error in studies of the productivity effects of multinational corporations. Stanley (2005) uses funnel graph and metaregression methods to test for publication bias in studies of the price elasticity of water, finding that the overall published price elasticity is nearly three times the corrected value. Finally, Doucouliagos (2005) notes a significant positive relationship between t statistics and sample sizes in studies of the effect of economic freedom on growth, again pointing to the asymmetry of the funnel graph and, consequently, publication bias. Moreover, he observes that the inflation of overall effect size due to publication bias is strongest in the early and late stages of the literature.[4]

In political science, Sigelman (1999) was the first to raise publication bias as a potential concern in the discipline, finding that 54.7% of null hypotheses were rejected at the 5% significance level in one volume of the *American Journal of Political Science*. Gerber et al.'s (2000) analysis of the voting mobilization literature finds a strong negative relationship between effect size and sample size, a pattern consistent with publication bias. Finally, in a study closely related the research reported here, Gerber and Malhotra (2006) find that in leading journals in political science there are far more barely statistically significant results than barely insignificant results than can be explained solely by chance.

Within sociology, there does not appear to be any recent research documenting the extent of publication bias. Early work by Wilson, Smoke, and Martin (1973) showed that a strong majority of the articles appearing in the *AJS*, the *ASR*, and *Social Forces* that used significance tests rejected the null hypothesis (80.3%). While this evidence, and similar evidence described earlier in other disciplines (e.g., Sigelman 1999; Sterling 1959),

is suggestive of a bias toward publishing statistically significant results regardless of research quality, it is also consistent with researchers estimating models that include variables known to be important predictors and reporting significance tests for these variables as well as researchers designing studies with large samples. Recent work by Leahey (2005) discusses how the use of statistical significance testing at arbitrary levels (and the ubiquitous use of "stars" to indicate significance) spread throughout sociology due to a contagion effect. The use of such hypothesis tests was not determined by suitability of the practice to the data, but instead by social factors such as their use by prestigious institutions and preferences of certain journal editors.

Considering that evidence of publication bias has been found throughout the social sciences since the 1950s, it is surprising that there is no extensive analysis of sociology research. Our article seeks to help remedy this omission by conducting an examination of multiple volumes of the leading journals over two periods. We examine the literatures using a caliper test, which we describe formally in the next section.

Using a Caliper Test to Detect Publication Bias

The statistical analysis in this article follows the approach used by Gerber and Malhotra (2006). This section discusses the rationale for their approach and describes its implementation. The caliper test detects publication bias by comparing the number of observations in equal-sized intervals just below and just above the threshold value for statistical significance. If there are an unusually large number of observations just over the critical value, this is taken as evidence of publication bias. If the imbalance is sufficiently great, the probability that 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. Keep in mind that the caliper test does not simply inform us of whether most published results are significant; we would expect this to be the case since researchers have priors and intuitions with respect to their stated hypotheses. Rather, the caliper test allows us to detect discontinuities around arbitrary significance levels, an unambiguous sign of bias. This section presents a brief formal discussion of the test.

The caliper test is based on the observed z scores. For any particular coefficient estimate, let $F(z)$ be the sampling distribution of the z score. Assume the sampling distribution F is continuous over the interval $[c, c'']$.

Under standard regularity conditions (see Greene 1996, sec. 6.7.3), the asymptotic sampling distribution of z corresponding to the least squares estimator is a continuous distribution; in particular, it is a normal distribution with a mean equal to the true value of the causal effect divided by the standard deviation of the estimate and a standard deviation equal to 1. The sampling distributions of the z scores produced by maximum likelihood estimation are also based on an asymptotically normal sampling distribution. For any number c, the probability that a draw from $F(z)$ is between c' and c, $c' > c$, is

$$F(c') - F(c). \tag{1}$$

The conditional probability that a draw from F is in the interval $[c, c']$, given F is drawn from $[c, c'']$, where $c'' > c' > c$, is

$$[F(c') - F(c)]/[F(c'') - F(c)]. \tag{2}$$

Since F is a continuous distribution, $F(x+e) - F(x) = (e)\,f(x) + E$, where $f(x)$ is the probability density function for z and E is an approximation error that equals zero if F is linear and approaches zero as e approaches zero for any continuous probability distribution. When the interval $[c, c'']$ is small and/or $f'(x)$ is small near c, the ratio in equation (2) can be approximated to first order by

$$[c' - c]\,f(c)/[c'' - c]\,f(c). \tag{3}$$

This approximation is exact if $f'(x) = 0$ over $[c, c'']$. Canceling the $f(c)$ terms, this result shows that for small intervals or when $f'(x) = 0$, the conditional probability of observing an outcome that falls in a subset in an interval equals the relative proportion of the subset to the interval. If $c' - c = c'' - c'$, then $[c', c'']$ is half the interval, and p is the conditional probability that an outcome falls in the upper subinterval, $p = .5$. For N independent draws in which the outcome falls in the $[c, c'']$ interval, the number of occurrences in a particular subset of the interval, where each occurrence in the subset has a probability p, is distributed binomial (N, p).

Concerns regarding publication bias focus on the effect of critical values on what is published. A natural choice for a statistical test for publication bias is therefore to set $c' = 1.96$, the critical value associated with the 5% significance level for a two-sided test, and examine a narrow interval of outcomes on either side of 1.96. Under the null hypothesis that what is published is a random draw from a sampling distribution, given that N occurrences fall within $1.96 - \varepsilon$ and $1.96 + \varepsilon$, the number of occurrences above 1.96 is distributed approximately binomially $(N, .5)$. The statistical

analysis in the next section evaluates whether the observed pattern of z scores is consistent with the null hypothesis or an alternative hypothesis, which states that there is an excess of occurrences above the critical value (that is, H_0: $p = .5$ vs. H_1: $p > .5$).

Before turning to the data, we discuss a few issues with the test. Although the empirical findings reported later in the article are stark and unlikely to be sensitive to the approximations used in generating the caliper test, it is still useful to reflect on the general robustness of the caliper test. The approximation used in equation (3) is likely to be quite accurate for small intervals above and below 1.96. This follows from the fact that the change in the height of the probability density function for the normal distribution is relatively modest over intervals of $\pm .2$ standard deviations, the interval size used in the most important tests reported later in the article.

This is illustrated by a few examples of how the approximation performs. For these examples, the caliper test compares the frequency of z-score estimates in the interval [1.8, 2] versus [2, 2.2]. First, suppose that the sampling distribution that is producing the z-score estimates is centered on 2. Based on the normal distribution, this implies that the probability of an event falling in each of the .2 standard deviation intervals is 7.9%, and the relative probability of falling in the upper interval is 1/2. This is exactly the $p = .5$ assumption used in the caliper test. Next, suppose the distribution is centered on 1 rather than 2. Since the distribution's mean is less than 2, the probability density of the normal distribution is thicker over the lower interval, [1.8, 2], than the upper interval, [2, 2.2]. The absolute probability of a draw in the lower portion of the caliper test range is 5.3%, and the relative probability is .55, not .5.[5] This is reversed if the distribution is centered at 3 rather than 1, in which case occurrences in the higher interval are slightly more, rather than slightly less, likely. While these values for p are not exactly .5, they are fairly close to .5.

In contrast to these relatively small deviations from 50-50, if the sampling distribution for z has a mean value much greater than the critical value 2 or much less than -2, the conditional probability of an occurrence in the lower versus upper portion of the partition can deviate from .5 by an appreciable amount. However, the probability that *any* draw from such a distribution is found in either the lower or upper partition is negligible. If the sampling distribution is centered on 5 (or -5), there is a 2:1 probability that a random draw will fall in the upper interval of the caliper test, but the chance of any draw between 1.8 and 2.2 is less than 1/500. As a result, distributions centered away from the critical value do not contribute materially to the statistical test.

Data

This section describes how we constructed the sample for our statistical analysis. The data set was extracted from the studies reported in three recent volumes of the *ASR*, the *AJS*, and *TSQ*. We included *TSQ* in our analysis because some may argue that the *ASR* and the *AJS* (widely considered the top two journals in sociology) specialize in those studies that are statistically significant, whereas the other studies are published, but in lesser journals.[6]

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. In this investigation, to determine 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 describe the procedures we used.

In conducting the caliper tests, we analyzed volumes 68 to 70 (2003 to 2005) from the *ASR*, volumes 109 to 111 (2003 to 2006) from the *AJS*, and volumes 44 to 46 (2003 to 2005) from *TSQ*. The total number of articles (excluding review essays, book reviews, and comments) yielded 118 articles from the *ASR*, 105 from the *AJS*, and 96 from *TSQ*.

Not all 319 articles could be tested for publication bias using the caliper method. Purely theoretical articles did not contain statistical tests. For those articles 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.[7]

We examined only articles that listed a set of hypotheses prior to presenting the statistical results.[8] Articles use slightly different conventions for naming hypotheses in this fashion (e.g., "Hypothesis 1," "H1," "H_1," "the first hypothesis"). Furthermore, some authors italicize or bold hypotheses within paragraphs while others indent them. Of the original 319 articles collected, 79 used this convention of explicitly listing hypotheses (36 in the *ASR*, 23 in the *AJS*, and 20 in *TSQ*). Thirteen articles were removed from the sample because they did not report standard errors or precise t statistics (4 in the *ASR* and the *AJS* and 5 in *TSQ*). We further restricted our sample to those articles that had 38 or fewer coefficients linked to the hypotheses. This appeared to be a natural cutoff as the article

Table 1
Summary of Exclusions During the Selection Process

	ASR	*AJS*	*TSQ*	Total
Total articles	118	105	96	319
Total that list hypotheses	36	23	20	79
Total that report standard errors	32	19	15	66
Selected articles ($\leq$38 coefficients)	20	14	12	46

Note: *AJS* = *American Journal of Sociology*; *ASR* = *American Sociological Review*; *TSQ* = *The Sociological Quarterly*.

with the next fewest coefficients had 48.[9] Applying this rule, we excluded 20 articles (12 in the *ASR*, 5 in the *AJS*, and 3 in *TSQ*). 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 an article 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 46 articles (20 from the *ASR*, 14 from the *AJS*, and 12 from *TSQ*) of the original 319 to subject to the caliper test. These selection steps are summarized in Table 1.

The next step is to link the explicitly stated hypotheses to specific regression coefficients. This could usually be done by visually inspecting regression tables, which often stated parenthetically the hypotheses associated with each coefficient. We confirmed that we were recording the correct figures by reading sections of the article and identifying which regression coefficients applied to which hypotheses. We remained agnostic about which regressions to include in the data set, including all specifications (full and partial) in the data analysis.[10] An example of how regression coefficients were selected is presented in the online appendix.

We also examined volumes 58 to 60 (1993 to 1995) from the *ASR* to see if the degree of bias has changed over time. The *AJS* and *TSQ* did not produce enough articles that met our criteria during this time period (8 and 1, respectively) for the purposes of conducting the caliper test. Out of 149 total articles published in the *ASR* during this time period, 35 used the convention of explicitly testing hypotheses, 24 fully reported standard errors or *t* statistics, and, of these, 15 did not exceed our cutoff for number of coefficients reported.

Because we examine only the tractable set of articles that explicitly list hypotheses, our results may not generalize to the full set of studies published in the journals. Since our article is about publication bias in the context of hypothesis testing, the subset of studies we examine is a reasonable place to initiate an investigation of sociological research. Articles in which scholars examine a large number of explanatory variables—usually in exploratory studies—without explicitly stating specific hypotheses raise a completely different and potentially interesting set of concerns.

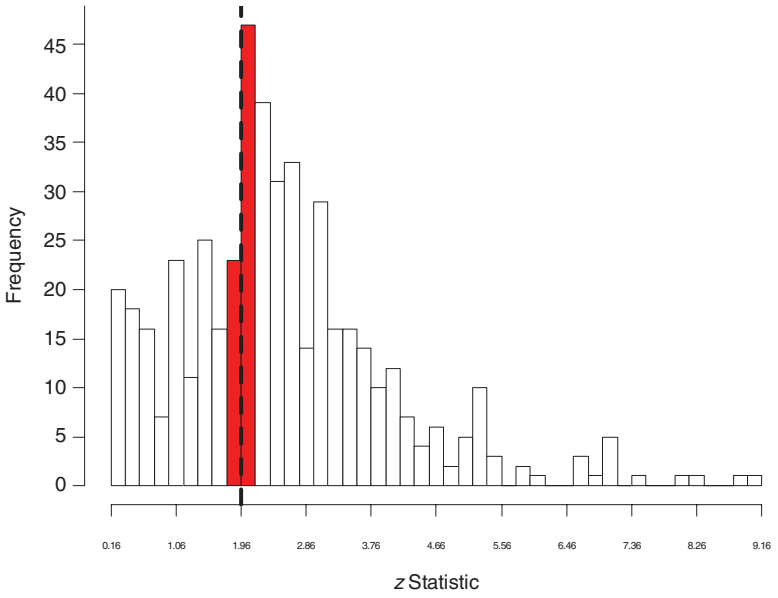
Results

In this section, we report the results of caliper tests, using calipers of several different sizes, for studies from the *ASR*, the *AJS*, and *TSQ* over the last few years. As described previously, caliper tests compare the number of coefficients with z scores within a specified interval above and below critical values. Before moving to these tests, it is useful to examine the relative frequency of z scores across a much broader range of values. Figures 1 and 2 show the distribution of z scores for coefficients reported in the three journals for two- and one-tailed hypotheses, respectively.[11] 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 three journals just over the critical value of 1.96 (see Figure 1). 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 2). These distributions show that there is greater density of published findings in the region barely above the arbitrary .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 both figures. Furthermore, the interval just below the critical value is close to being the local minimum in the one-tailed case. In other words, it is one of the smallest bins in its neighborhood.

Whereas the figures use a 10% caliper, Table 2 shows the results using a variety of calipers, with one- and two-sided tests pooled. The imbalance of findings seen in the histograms is robust across caliper sizes, journals, and publication dates. However, there is a pattern in the results. 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

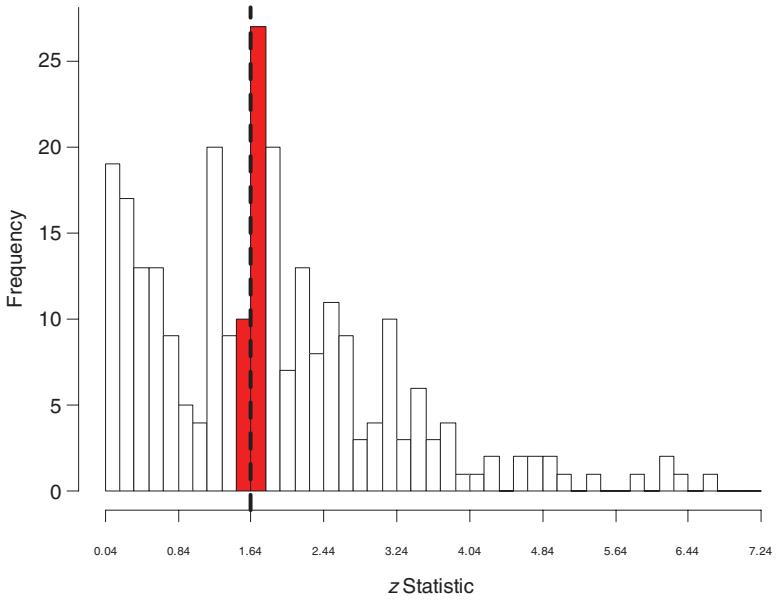
Figure 1
Histogram of *z* Statistics From the *American Sociological Review*, the *American Journal of Sociology*, and *The Sociological Quarterly* (Two-Tailed)



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 nearly 4:1 for the combined results for the 5% caliper, the ratio is between 2:1 and 3:1 for the wider calipers. Indeed, using a tiny 2.5% caliper for the *ASR*, we observe 10 results above the caliper and 2 results below, a ratio of 5:1. For the *AJS*, there are 12 results above the 2.5% caliper and 2 results below it. And in *TSQ*, the ratio is 10:3. The fact that the results are most dramatic in the narrow regions nearest to the critical value provides additional evidence that 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

Figure 2
Histogram of *z* Statistics From the *American Sociological Review*, the *American Journal of Sociology*, and *The Sociological Quarterly* (One-Tailed)



independent, we calculated the probabilities that imbalances of the magnitude we observe would occur by chance. We find that it is easy to reject the hypothesis that the observed patterns are due to chance for the data overall, for each of the three journals, for one-sided tests, and for two-sided tests. For instance, for the 10% caliper for the three journals combined, the chance that we would observe the imbalance we do by chance alone is less than 1 in 15,000. For the 5% caliper, the chance is about 1 in 100,000. It is also striking how similar the ratios are across the three journals. For the 5% caliper, the ratios vary from 3.25:1 to 4:1, and for the 15% caliper, they vary from about 2.4:1 to 2.9:1.

There is little evidence that the *ASR* and the *AJS* specialize in statistically significant findings while *TSQ* picks up those meritorious papers rejected for falling just short of statistical significance. The caliper test

Table 2
Caliper Tests of Publication Bias in the *ASR*, the *AJS*, and *TSQ*

	Over Caliper	Under Caliper	*p* Value[a]
ASR (Vols. 68-70)			
5% caliper	15	4	.01
10% caliper	26	15	.06
15% caliper	47	17	<.001
20% caliper	54	19	<.001
AJS (Vols. 109-111)			
5% caliper	16	4	.006
10% caliper	25	11	.01
15% caliper	41	14	<.001
20% caliper	48	18	<.001
TSQ (Vols. 44-46)			
5% caliper	13	4	.02
10% caliper	22	7	.004
15% caliper	26	11	.01
20% caliper	30	20	.10
Combined (recent vols.)			
5% caliper	44	12	<.001
10% caliper	73	33	<.001
15% caliper	114	42	<.001
20% caliper	132	57	<.001
ASR (Vols. 58-60)			
5% caliper	17	2	<.001
10% caliper	22	5	<.001
15% caliper	27	11	.007
20% caliper	30	15	.02

Note: "Over caliper" indicates the number of results that are between 0 and *X*% greater than the 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 to 1.81 for one-tailed tests and 1.96 to 2.16 for two-tailed tests. "Under caliper" represents the number of results that are between 0 and *X*% less than the critical value. For the 10% caliper, the under-caliper range is about 1.48 to 1.64 (1.76-1.96). *AJS* = *American Journal of Sociology*; *ASR* = *American Sociological Review*; *TSQ* = *The Sociological Quarterly*.
a. Based on density of binomial distribution (one-tailed).

shows very similar patterns for *TSQ* as for the *ASR* and the *AJS*; the results for the 10% caliper are even more extreme for *TSQ* than the other journals.

Examining the older volumes of the *ASR*, we find that the degree of imbalance just above versus just below the critical value is not decreasing

over time. As shown in the bottom panel of Table 2, the imbalance of *t* statistics above and below the calipers is not substantially different between the mid-1990s and the mid-2000s. Whereas the imbalance in the narrow calipers is more pronounced for the earlier volumes of the *ASR*, the most recent volumes exhibit more lopsidedness in the wider calipers.[12] Combining the data from the recent and older journals, under the assumption of independence the probability of observing by chance the pattern produced by the 5% caliper is less than 1 in 10 million and the chance of observing the pattern produced by the 10% caliper is less than 1 in 1 million.

The statistical tests we perform may slightly overstate the rarity of the patterns we report under the null hypothesis. One factor that complicates the analysis is that the studies included in our data set often contributed more than one coefficient. This suggests each coefficient cannot be viewed as statistically independent, though we have no way to determine the covariances from the information contained in the articles and accounting for statistical dependence in our analysis would be complex. In calculating *p* values, we assume that the coefficients can be treated as independent. This will produce spurious results if, for example, the fact than one coefficient from a hypothesis test in a study has a *z* score of between 2 and 2.2 materially increases the relative probability that another coefficient from a hypothesis in the study has a *z* score between 2 and 2.2 rather than between 1.8 and 2, something unlikely for a continuous distribution. While we have not undertaken the exercise, perhaps pathological cases can be constructed. We take some comfort, however, in the similarity of the findings across a range of independent tests presented in Table 2 and elsewhere.

One further check of the results is to restrict attention to those studies that contributed only a couple of coefficients to the statistical test. Table 3 shows the results for the *ASR*, the *AJS*, and *TSQ* broken down by the number of coefficients per study that fell within the range associated with various caliper values. Looking only at the studies that contributed one or two coefficients shows the same 3:1 imbalance reported earlier—17:6 using the 10% caliper. The odds of an imbalance as great as or greater than this is, under the null hypothesis of equal probability, less than .02.

The results in Tables 2 and 3 show substantial evidence that the .05 significance level is affecting published research. There are situations where it might be expected that publication bias will be more rather than less pronounced. For instance, publication bias might be more frequently observed in the early stages of a literature when a key explanatory variable

Table 3
Caliper Tests of the *ASR*, the *AJS*, and *TSQ* by Coefficients Contributed per Study

	5% Caliper			10% Caliper			15% Caliper			20% Caliper		
Coefficients Contributed	Over	Under	Studies	Over	Under	Studies	Over	Under	Studies	Over	Under	Studies
1	7	4	11	7	6	13	4	3	7	5	3	8
2	14	2	8	10	0	5	14	2	8	10	4	7
3	5	1	2	11	4	5	9	6	5	10	2	4
4	7	1	2	12	8	5	18	6	6	14	6	5
5	11	4	3	15	5	4	7	3	2	15	10	5
6	0	0	0	14	4	3	22	8	5	11	1	2
7	0	0	0	0	0	0	10	4	2	24	11	5
8	0	0	0	0	0	0	14	2	2	11	5	2
9	0	0	0	0	0	0	7	2	1	7	2	1
10	0	0	0	4	6	1	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	15	7	2
12	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	9	6	1	0	0	0
16	0	0	0	0	0	0	0	0	0	10	6	1
Total	44	12	26	73	33	36	114	42	39	132	57	42

Note: "Coefficients contributed" refers to the number of results for a given study within a certain caliper. For instance, one study had 15 coefficients within the 15% caliper—9 over it and 6 below it. *AJS* = *American Journal of Sociology*; *ASR* = *American Sociological Review*; *TSQ* = *The Sociological Quarterly*.

is being offered to the discipline. Conversely, in the later stages of a literature, the novel finding is to show *no* significant relationship between the explanatory variable and the phenomenon of interest, thereby debunking a previously established relationship. Additionally, there may be less of a need to increase the z statistic of a barely insignificant result if a study has other statistically strong results in it as well. We find little evidence of this in the Table 3 data, which shows no clear relationship between the number of coefficients related to the hypotheses and the degree of imbalance in results near critical values. However, a more detailed study of this question might find support for this very plausible conjecture. Other research practices may be the consequence of pressure to breach conventional significance levels. For example, a regression model can be estimated, with the significant results circled, which are then used to formulate the main hypotheses of the article. In this case, explanatory variables are presented not as the consequence of theory, but rather as products of data mining.

Discussion

We assert that publication bias is responsible for the results shown in Figures 1 and 2 and Table 2. Are there any other 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 unlikely 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. It is possible that some researchers are using adaptive sampling, deciding to get more data if their results are nearly, but not quite, statistically significant. If so, the significance levels used for the statistical tests reported are incorrect. Furthermore, if this practice is combined with a file-drawer problem, in which papers that are not boosted over the statistical significance threshold by additional sampling never appear, biased literatures will result. The findings of Gerber et al. (2000) suggest that in fact the ''failures'' do not find their way into print.

Perhaps the true effects are arrayed in a manner that produces the imbalance observed in the data? As explained previously, this is highly unlikely. If the true effects happen to fall such that the sampling distributions from which the z scores are drawn are centered on 2, then the

probability mass in adjacent intervals of .2 near the critical values is nearly identical (the probability of a draw from a standard normal being between 0 and .2 is about 7.5%, while between .2 and .4 it is about 7.1%). The ratio of probability mass in adjacent intervals gets further from 1 when you move away from the center of the normal. However, the ratio would deviate from 1 an appreciable amount only when the true effects are very large. In this case the relative probability of a draw a bit greater than 1.96, for example, is noticeably higher than the event of a draw less than 1.96. But since *either* of these events will be very rare when the true effect is very large, these cases will not have a material effect on our findings.[13]

There is convincing evidence that something is causing a distortion of the published literature around the critical values for statistical significance. How important is this? If the extent of the matter is a small distortion in the reported results, then publication bias is unfortunate but not catastrophic. If it produces a 5% or 10% change in reported z scores, eliminating this problem is an accomplishment of the order associated with justly celebrated methodological innovations such as the use of robust standard errors.[14]

However, this somewhat sanguine view is probably not justified. First, for hypothesis testing, the effect of publication bias may be far greater. If a sizable portion of the outcomes that fall in the range of 1.6 to 2 are either not submitted, are not published, or are reworked, then the chance of a Type I error is really much greater than that which is reported. For example, under the null hypothesis that the true effect is zero, if two thirds of the time a result with an absolute value between 1.6 and 2 gets somehow moved to something greater than 2, a test with a Type I error rate of 5% really has a Type I error rate of approximately 9%, which is nearly twice the reported rate.

Second, we have focused on only one consequence 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. The professional incentives governing what is deemed important and publishable may vary over the life cycle of a literature, and this may cause publication bias. In the germinal stages of a literature's development, the incentive might be to find previously undiscovered statistically significant results. However, once that literature has become established, the incentive may shift to discrediting these initial findings by reporting statistically weak relationships between relevant variables. Publication bias would function differently in the early stage as opposed to the later stage, suggesting that

the amount of distortion of estimates due to the critical values is underestimated by our analysis. An exploration of this phenomenon would further shed light on the mechanisms of publication bias and is a worthy topic of future research.

Third, the clear evidence of the presence of one form of publication bias makes it plausible that there may be other forms of publication bias as well, unrelated to the critical values. Future research can seek to identify the sources of this bias. For example, a more intensive analysis of the use of multiple specifications may identify how variables are added and removed from regression analyses or how the sample is segmented in various ways to push results into the territory of statistical significance.

Fourth, our method of selecting coefficients entailed examining the hypotheses the authors themselves highlighted within their articles. However, it is possible that a subset of these could be considered the most critical results, and the imbalance may be even greater when focusing on these key hypotheses. Identifying such coefficients would involve substantial practical difficulties and subjective coding judgments on our part and is therefore beyond the scope of the current article. Nevertheless, replicating our analyses by identifying the one or two key findings of each article represents a fruitful avenue for future research.

Our evidence cannot definitively adjudicate between several non–mutually exclusive explanations for the sources of publication bias. On one hand, it could be that researchers submit barely insignificant findings, but that journals reject them. On the other hand, barely insignificant findings may find their way into file drawers. However, as seen in Figure 2, there are many reported insignificant (but not barely insignificant) findings, suggesting that neither the "publication" nor "submission" explanations can be the whole story. Because articles tested, on average, 17.3 coefficients, it is possible that a mixture of significant and insignificant findings is sufficient for submission and publication. Finally, it is possible that researchers may be adjusting regression specifications and bifurcating samples to move results that are below the caliper above the caliper. However, if this were the sole factor driving the results, then we would have observed a sharp increase in insignificant findings in the bins beneath the lower bound of the caliper. While the distributions (particularly of the one-tailed tests presented in Figure 2) are suggestive of this pattern, the results are not conclusive. The next step forward in exploring the mechanisms underlying publication bias is to examine the pool of papers submitted to journals, a project we hope to undertake.

The goal of this article is to raise awareness of publication bias in sociology. We have found that many more results are published just over the $p = .05$ threshold than below it, implying a certain amount of collective bias in parameter estimates. Our results suggest that as reviewers, editors, and researchers, sociologists appear to be far too conscious of the .05 significance level and that this might cause important distortions in how knowledge advances in sociology.

Several steps might be considered to reduce publication bias. If it is thought that the imbalance in results just over the statistical significance threshold is due to specification searching and arbitrary choices, it would be useful to have a fresh pair of eyes look over these implementation decisions and detect when they are material to the results. This kind of replication work can already be done when scholars provide their data to the research community, but the efforts are ad hoc and, perhaps because the initiative to audit is taken by the individual scholar, often disputatious. Perhaps a better system would be for one or two articles in each issue of a journal to be randomly selected for communal audits. The journal could announce which article or articles in the issue have been randomly selected for audit. The data and programs would be made available and informative comments on the selected article would be published in the next issue, and all comments would be posted on a Web site. This would establish a tournament with clear rewards and therefore incentives for researchers to do the checking as well as depersonalize the process of selecting an article to replicate. The idea would not be to humiliate scholars publicly, but rather to encourage researchers to explore the robustness of their results to minor variations in specification choice *prior* to submission and to enhance the reputations of researchers whose work is found to be robust.

Another institutional reform that might limit the number of "lost" studies, limit specification searching, and reduce ex post identification of subgroups, outcomes, or hypothesis tests is the establishment of study registries. Prior to conducting research, a description of the proposed research would be filed with a central registry. 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 in medical clinical trials and should be considered for sociological experimentation as well, we propose that registries can be usefully extended to observational studies that involve the analysis of data not yet available to the researcher, such as survey or election results.

The evidence presented here demonstrating that critical values affect the set of published articles should encourage greater emphasis on confidence intervals rather than point estimates. Statistically meaningless movements in estimation results from just above critical values to just below may produce a severe reaction in researchers and reviewers. However, presenting the same small movement in the results as a shift in a confidence interval would provide a more intuitive representation of the nonevent such a change typically is. Additionally, the $X\%$ classical confidence interval often has the added advantage of being interpretable in a Bayesian framework as a set of values that captures $X\%$ of the posterior distribution (see Gelman et al. 2004, chap. 4).

In recent years there has been impressive progress in empirical research in sociology. The leading journals are filled with articles applying sophisticated statistical methods to important empirical questions. As the quality of individual studies improves, this should translate into increasingly reliable knowledge. However, publication bias interferes with progress toward this goal, and our article provides substantial evidence of publication bias in the discipline's leading journals. Addressing the full range of questions raised by this finding is an important new challenge for those concerned with, and those who work to improve, the accuracy of the discipline's quantitative research.

Notes

1. The submission instructions of the *American Sociological Review* (*ASR*) read: "Use asterisks . . . to indicate significance at the $p < .05$, $p < .01$, and $p < .001$ levels. . . . Generally, results at $p > .05$ (such as $p < .10$) should not be indicated as significant." Such reporting conventions may have the unintended consequence of encouraging the practice of adjusting results to achieve certain significance levels.

2. To correct for potential errors-in-variables bias due to the fact that precision is often itself an estimate, Stanley (2005) argues that sample size can be used in place of or as an instrument for the standard error to measure precision.

3. There also exist several techniques to model selection and correct for publication bias using weighted distribution theory (Dear and Begg 1992; Hedges 1984, 1992; Iyengar and Greenhouse 1988), Bayesian statistics (Cleary and Casella 1997), and maximum likelihood (Copas 1999). In addition to selection models, sensitivity analyses are often used to ascertain the severity of the bias (Duval and Tweedie 2000; Gleser and Olkin 1996; Rosenthal 1979).

4. One exception is Doucouliagos, Laroche, and Stanley's (2005) examination of the union productivity literature, which is not found to suffer from publication bias. The funnel graph is nearly perfectly symmetrical, implying no structural relationship between effect size and precision.

5. This calculation, as well as the previous calculation, ignores the very small probability that the outcome falls 3 or 4 standard deviations from the mean and appears in the caliper test intervals for the z score associated with draws from the normal in either $[-1.8, -2]$ or $[-2.2, -2]$.

6. According to Michael Allen's "core influence" measure (Allen, 2003) the *American Journal of Sociology* (*AJS*) is the most influential sociology journal, the *ASR* is the 2nd most influential, and *The Sociological Quarterly* (*TSQ*) is the 10th most influential.

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

8. The vast majority of these hypotheses required a statistically significant result to be accepted. The handful of articles that had a hypothesis predicting a null finding did not contribute any coefficients in any caliper tested. Hence, these "null hypotheses" do not affect the results presented below.

9. Of course there is nothing methodologically incorrect about testing many hypotheses. Some articles had more than 100 coefficients that tested hypotheses. For example, Elman and O'Rand (2004, *AJS*) listed five sets of hypotheses (each represented by 4–16 variables) tested across six specifications. This produced 106 coefficient estimates.

10. In some cases, the reported t statistic (estimate divided by standard error) was inconsistent with the notation of significance (i.e., presence of an asterisk) 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.

11. Very large outlier z scores are omitted to make the x-axis labels readable. The omitted cases are a small percentage (7.0% for both the one-tailed and two-tailed tests) of the sample and do not affect the caliper tests.

12. As mentioned above, there were not enough articles fulfilling our criteria from the *AJS* and *TSQ* during this time period to adequately conduct the caliper test. For the articles we did identify in the *AJS* for the 10% caliper, we observed the same pattern with the limited data that were available. There were five coefficients above the caliper and two coefficients below the caliper. For *TSQ*, there was one coefficient above the caliper and zero coefficients below the caliper.

13. Examining the empirical distribution of z scores (Figures 1 and 2) it appears that the contribution to the caliper test of sampling distributions centered on values less than 2 equals or exceeds that coming from distributions centered on values greater than 2. This suggests the caliper test is conservative.

14. One important difference, however, is that in contrast with incorrect standard errors, problems with persistent bias are not eliminated asymptotically.

References

Allen, Michael P. 2003. "The 'Core Influence' of Journals in Sociology Revisited." Footnotes. 31:7-10.

Allison, D. B., M. S. Faith, and B. S. Gorman. 1996. "Publication Bias in Obesity Treatment Trials?" *International Journal of Obesity* 20:931-37.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, With Tests for Publication Bias." *Labour Economics* 6:453-70.

Begg, Colin B. and Jesse A. Berlin. 1988. "Publication Bias: A Problem in Interpreting Medical Data." *Journal of the Royal Statistical Society, Series A* 151:419-63.

Berlin, Jesse A., Colin B. Begg, and Thomas A. Louis. 1989. "An Assessment of Publication Bias Using a Sample of Published Clinical Trials." *Journal of the American Statistical Association* 84:381-92.

Campbell, Donald T. 1969. "Reforms as Experiments." *American Psychologist* 24:407-29.

Card, David and Alan B. Krueger. 1995. "Time-Series Minimum Wage Studies: A Meta-Analysis." *American Economic Review* 85:238-43.

Cleary, Richard J. and George Casella. 1997. "An Application of Gibbs Sampling to Estimation in Meta-Analysis: Accounting for Publication Bias." *Journal of Educational and Behavioral Statistics* 22:141-54.

Copas, John. 1999. "What Works?: Selectivity Models and Meta-Analysis." *Journal of the Royal Statistical Society, Series A* 161:95-105.

Correll, Shelley J. 2001. "Gender and the Career Choice Process: The Role of Biased Self-Assessments." *American Journal of Sociology* 106:1691-730.

Coursol, Allan and Edwin E. Wagner. 1986. "Effect of Positive Findings on Submission and Acceptance Rates: A Note on Meta-Analysis Bias." *Professional Psychology* 17:136-37.

Dear, Keith B. G. and Colin B. Begg. 1992. "An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis." *Statistical Science* 7:237-45.

De Long, J. Bradford and Kevin Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100:1257-72.

DiPrete, Thomas A. and Henriette Engelhart. 2004. "Estimating Causal Effects With Matching Methods in the Presence and Absence of Bias Cancellation." *Sociological Methods & Research* 32:501-28.

Doucouliagos, Chris. 2005. "Publication Bias in the Economic Freedom and Economic Growth Literature." *Journal of Economic Surveys* 19:367-87.

Doucouliagos, Chris, Patrice Laroche, and T. D. Stanley. 2005. "Publication Bias in Union-Productivity Research?" *Industrial Relations* 60:320-47.

Duval, Sue and Richard Tweedie. 2000. "A Nonparametric 'Trim and Fill' Method of Accounting for Publication Bias in Meta-Analysis." *Journal of the American Statistical Association* 95:89-98.

Easterbrook, P. J., J. C. Keruly, T. Creagh-Kirk, D. D. Richman, R. E. Chaisson, 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* 266:2713-18.

Egger, Matthias, George D. Smith, Mariin Schneider, and Christoph Minder. 1997. "Bias in Meta-Analysis Detected by a Simple, Graphical Test." *British Medical Journal* 315:629-34.

Elman, Cheryl and Angela M. O'Rand. 2004. "The Race Is to the Swift: Socioeconomic Origins, Adult Education, and Wage Attainment." *American Journal of Sociology* 110:123-60.

Epstein, William M. 2000. "Confirmational Response Bias Among Social Work Journals." *Science, Technology, & Human Values* 15:9-38.

———. 2004. "Confirmational Response Bias and the Quality of the Editorial Processes Among American Social Work Journals." *Research on Social Work Practice* 14:450-58.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. 2d ed. Boca Raton, FL: Chapman & Hall.

Gerber, Alan S., Donald P. Green, and David Nickerson. 2000. "Testing for Publication Bias in Political Science." *Political Analysis* 9:385-92.

Gerber, Alan S. and Neil Malhotra. 2006. "Can Political Science Literatures Be Believed? A Study of Publication Bias in the *APSR* and the *AJPS*." Presented at the annual meeting of the Midwest Political Science Association, April 20-23, Chicago, IL.

Glass, G. V., B. McGaw, and M. L. Smith. 1981. *Meta Analysis in Social Research*. Beverly Hills, CA: Sage.

Gleser, L. J. and I. Olkin. 1996. "Models for Estimating the Number of Unpublished Studies." *Statistics in Medicine* 15:2493-507.

Gorg, Holger and Eric Strobl. 2001. "Multinational Companies and Productivity: A Meta-Analysis. *Economic Journal* 111:723-39.

Greene, William H. 1996. *Econometric Analysis*. 3d ed. Upper Saddle River, NJ: Prentice Hall.

Greenwald, Anthony G. 1975. "Consequences of Prejudice Against the Null Hypothesis." *Psychological Bulletin* 82:1-20.

Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory Into Practice." *Annual Review of Sociology* 30:507-44.

Hedges, Larry V. 1984. "Estimation of Effect Size Under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences." *Journal of Educational Statistics* 9:61-85.

———. 1992. "Modeling Publication Selection Effects in Meta-Analysis." *Statistical Science* 7:246-55.

Iyengar, Satish and Joel B. Greenhouse. 1988. "Selection Models and the File Drawer Problem." *Statistical Science* 3:109-35.

Leahey, Erin. 2005. "Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology." *Social Forces* 84:1-23.

Levois, M. E. and M. W. Layard. 1995. "Publication Bias in the Environmental Tobacco Smoke/Coronary Heart Disease Epidemiologic Literature." *Regulatory Toxicology and Pharmacology* 21:184-91.

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

McLeod, B. D. and J. R. Weisz. 2004. "Using Dissertations to Examine Potential Bias in Child and Adolescent Clinical Trials." *Journal of Consulting and Clinical Psychology* 2:235-51.

Mouw, Ted. 2006. "Estimating the Causal Effect of Social Capital: A Review of Recent Research." *Annual Review of Sociology* 32:79-102.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108:937-75.

Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86:638-41.

Sigelman, Lee. 1999. "Publication Bias Reconsidered." *Political Analysis* 8:201-10.

Simes, R. J. 1986. "Publication Bias: The Case for an International Registry of Clinical Trials." *Journal of Clinical Oncology* 4:1529-41.

Stanley, T. D. 2005. "Beyond Publication Bias." *Journal of Economic Surveys* 19:309-37.

Stanley, T. D. and Stephen B. Jarrell. 1989. "Meta-Regression Analysis: A Quantitative Method of Literature Surveys." *Journal of Economic Surveys* 3:161-70.

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.

Sterling, T. D., W. L. Rosenbaum, and J. J. Weinkam. 1995. "Publication Bias Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa." *American Statistician* 49:108-12.

Sutton, A. J., F. Song, S. M. Gilbody, and K. R. Abrams. 2000. "Modelling Publication Bias in Meta-Analysis: A Review." *Statistical Methods in Medical Research* 9:421-45.

Wilson, Franklin D., Gale L. Smoke, and J. David Martin. 1973. "The Replication Problem in Sociology: A Report and a Suggestion." *Sociological Inquiry* 43:141-49.

Alan S. Gerber is a 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 the Melvin & Joan Lane Graduate Fellow and a PhD candidate in the Department of Political Science at Stanford University. His research has been published in the *American Political Science Review* and the *Journal of Politics*. In July 2008, he will join the Stanford Graduate School of Business as an assistant professor of political economy.