JOURNAL OF
EMPIRICAL LEGAL STUDIES

Journal of Empirical Legal Studies
Volume 10, Issue 4, 603–611, December 2013

Foreword: Conference Bias

*Daniel E. Ho**

I. Overview

This annual special issue of the *Journal of Empirical Legal Studies* (*JELS*) celebrates the Seventh Annual Conference on Empirical Legal Studies (CELS), held at Stanford Law School on November 9–10, 2012.

There is much to celebrate. CELS received nearly 360 submissions for peer review, resulting in 103 paper presentations (each with individual discussants) and 24 poster presentations. Some 340 individuals from across the globe attended the conference. Indeed, empirical work has left few fields of law untouched, as the left panel of Figure 1, plotting submissions by field, demonstrates. Of course, some areas are subject to greater empirical scrutiny. Corporate governance and finance, criminal justice, law and psychology, and law and politics, in particular, drew the largest number of submissions, reflecting the interdisciplinary nature of empirical work pertaining to law. The right panel of Figure 1 plots the acceptance rate by topic—while there is some variability, the rates are comparable across fields (p value = 0.66).[1] And, as highlighted by the *Stanford Law Review* issue on the "Empirical Revolution in Law,"[2] this empirical movement in law is distinguished by its pervasiveness, engaging a wide range of scholars who do not necessarily produce primary empirical work themselves.

This issue is a capstone to CELS, publishing a select group of papers initially presented at the conference. Just how *select* are they? And do the "best" or published papers,

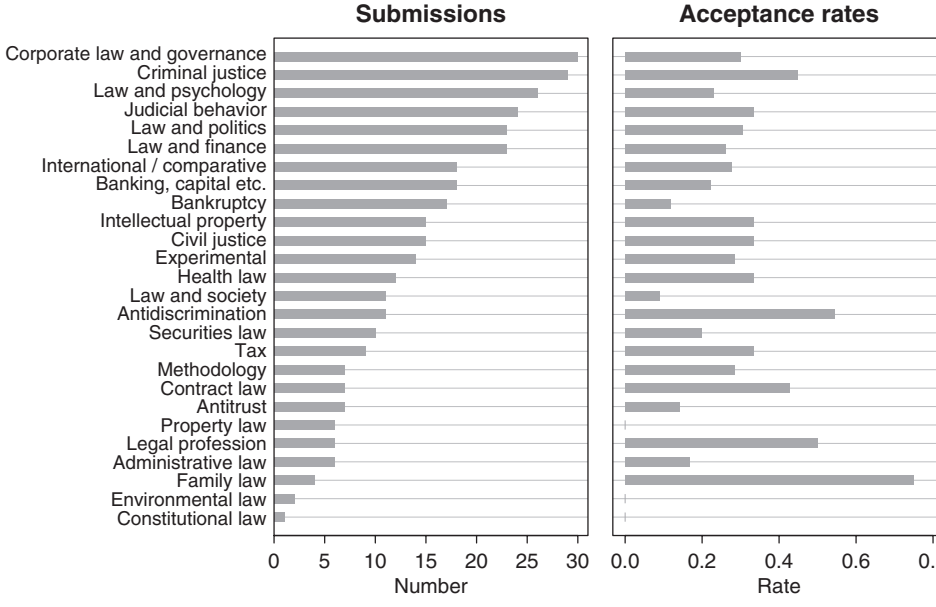
*Professor of Law & Robert E. Paradise Faculty Fellow for Excellence in Teaching and Research, Stanford Law School, 559 Nathan Abbott Way, Stanford, CA 94305; email: dho@law.stanford.edu; URL: http://dho.stanford.edu.

Thanks to Patrick Leahy for outstanding research assistance on this foreword and Dawn Chutkow, Ted Eisenberg, and Michael Morse for helpful comments. CELS would not have been possible without the extraordinary support and help of Deans Larry Kramer and Liz Magill, the Executive Director of the Society for Empirical Legal Studies (SELS), Dawn Chutkow, the Stanford Program Committee (Rob Daines, Deborah Hensler, Dan Kessler, Mike Klausner, Alison Morantz, and, especially, the indefatigable David Freeman Engstrom), the Stanford Program Group (Jackie Del Barrio, Trish Gertridge, Erin Lee, and Cassey Limgenco), the SELS Board of Directors (David Abrams, Jennifer Arlen, Bernie Black, Shari Diamond, John Donohue, Ted Eisenberg, Valerie Hans, Michael Heise, Geoff Miller, Eric Talley, and Emerson Tiller), the many Stanford-affiliated faculty members who refereed papers, and the corps of generous discussants at the conference.

[1]Because χ^2 tests are invalid for contingency tables with small cell counts (Cochran 1952), we calculate this p value via Monte Carlo simulation, drawing 1,000 contingency tables given fixed row and column totals (Patefield 1981). As a test statistic, we employ the maximum squared deviation from expected counts.

[2]65 Stan. L. Rev. 1195 (2013).

Figure 1: Distribution of papers at CELS 2012.



NOTE: The left panel plots the distribution of submissions by topic, based on consolidated SSRN author self-classifications. The right panel plots the acceptance rate. Four unclassified papers are omitted.

accumulated over the years, accurately represent our empirical knowledge about the legal system?

While there is much to applaud, this foreword sounds a note of caution about the practice of empirical work in law. Collecting data from all *JELS* articles published since its inception in 2004, as well as papers presented at the conference in 2012, we document considerable evidence of "publication bias": because of arbitrary statistical significance thresholds, published results may not represent true effects. Contrary to certain conventional notions of publication bias, however, we demonstrate that the phenomenon does not stem from publication per se. Bias appears, if anything, worse at the conference. *JELS* in fact appears *less* susceptible to bias in large part because of a greater focus on descriptive research and, possibly, an editorial process that mitigates particular forms of specification searching.

II. Publication Bias

To what extent does publication bias plague law? Although publication bias has been documented in a wide range of disciplines,[3] to our knowledge no systematic empirical

[3]See, e.g., Gerber and Malhotra (2008a) (political science), Gerber and Malhotra (2008b) (sociology), Sterling (1959), Masicampo and Lalande (2012) (psychology), Card and Krueger (1995), Brodeur et al. (2012) (economics), and Easterbrook et al. (1991) (medicine).

inquiry of it exists in law.[4] To test for the presence of publication bias in empirical legal studies, we apply the approach of Gerber and Malhotra (2008b) and Brodeur et al. (2012), which focuses on one observable manifestation: the densities of test statistics around a critical value (e.g., *z* scores around the critical value of 1.96 at $\alpha = 0.05$). In the absence of publication bias, we should not expect to see sharp discontinuities around the threshold.[5]

Our study also deepens our understanding of publication bias in three other ways. While it has been documented in other fields, much less is known about the precise mechanism generating publication bias. One conventional view is that journal editors and reviewers accept papers based on a simple cut-off rule, publishing findings only if they meet conventional significance levels (e.g., $\alpha = 0.05$) (see, e.g., Gigerenzer 2004:598–99). An alternative (but certainly not exclusive) mechanism centers on researchers who may opt not to submit research with statistically insignificant findings, perhaps in anticipation of a negative journal decision (the so-called file drawer problem) (see Rosenthal 1979). More pathologically, researchers may engage in specification searching until a hypothesis is rejected (see Leamer 1983). Our study sheds insight into these mechanisms by examining papers from two different stages of the research process.[6] Paradoxically, we find that "conference bias" may be worse than publication bias, suggesting that the practice is more endemic than an editorial rule.[7]

Second, we assess what kind of empirical work (experimental, observational, or descriptive) appears most susceptible to publication bias. Results from other fields have been varied. In medicine, Easterbrook et al. (1991) find publication bias more prevalent in observational studies than in clinical trials. In economics, Roth (1994) argues that Leamer's (1983) influential critique of econometric practices applies similarly to experimental approaches. In psychology, Gigerenzer (2004) argues that hypothesis testing undermines high-quality descriptive and exploratory statistics.

Third, our findings inform specific proposals to address publication bias. Some scholars advocate the inclusion of conferences and electronic paper repositories as a way toward reducing publication bias (see, e.g., Blumenthal 2007; Callaham & Wears 1998; Petticrew et al. 1999; Song 1999). The logic is simple: if null results are harder to publish formally in a journal, researchers should look to conference, electronic, and unpublished

---

[4]Publication bias has, of course, not gone entirely unmentioned in scholarship pertaining to law. Blumenthal (2007) and Pfaff (2010) discuss publication bias in the context of conducting meta-analyses of empirical legal research. Donohue and Wolfers (2006) adjust for publication bias in the context of research on capital punishment. Sporer and Goodman-Delahunty (2011) discuss the role of Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993), in meta-analyses for law and psychology.

[5]This approach is, of course, not the only way to assess publication bias. When examining studies of the same quantity of interest, for instance, a funnel plot provides a useful method to assess publication bias.

[6]See Gerber and Malhotra (2008a:321–22) (noting that "future research could determine the extent to which bias is due to selection on the part of reviewers and editors"). In the same spirit of the analysis here, Gerber and Malhotra proposed (unsuccessfully) to compare a sample of submissions to published articles.

[7]Of course, researcher decisions to file studies away or to engage in specification searching are themselves likely driven by journal cut-off rules.

studies to obtain unbiased estimates of population effects. Consistent with this logic, Glass et al. find effect sizes are larger in psychology journals than in unpublished works (1981: 66, tab. 3.4). Petticrew et al. (1999) find that medical conference proceedings with uncertain results were less likely to be published and argue that publication of conference proceedings would allay publication bias. Hopewell et al. (2005) advocate including "grey literature" (i.e., literature "not controlled by commercial publishers"), such as conference proceedings, in systematic reviews, an approach also proposed for law (see Blumenthal 2007). Our findings suggest that solutions focusing on such grey literature will be limited, at least in law.

III. Data Collection

To investigate the extent of publication bias, we developed a protocol to collect test statistics from all articles published in *JELS* since its inception (259 articles from 2004–2013) as well as all paper and poster presentations from CELS 2012 (127 papers). We classified eligible papers—those with formal statistical tests of association or presentation of statistical uncertainty—into three types: (1) *experimental* papers that employed randomization to study the causal effect of an intervention; (2) *observational studies* that focused on one or several primary inferences of interest (including causal inferences, but also descriptive inferences when such inferences were limited to specific hypotheses);[8] and (3) *descriptive studies* without particular primary inferences of interest (e.g., studies regressing an outcome on many explanatory variables without focused, specific hypotheses).

We then collected test statistics corresponding to the principal results from each eligible paper. Most commonly, these test statistics were regression coefficients with standard errors, t statistics, or p values. Experimental papers and observational studies naturally limited the set of test statistics from each paper. For descriptive studies, when there was no primary inference of interest, we collected test statistics of all coefficients from the principal model reported. To ensure that descriptive studies with many test statistics do not drive our findings, our analyses weight papers equally (by weighting test statistics inversely to the number of test statistics collected per paper). In all, we collect 3,709 test statistics, which we convert to z scores, whenever sufficient information is provided.[9]

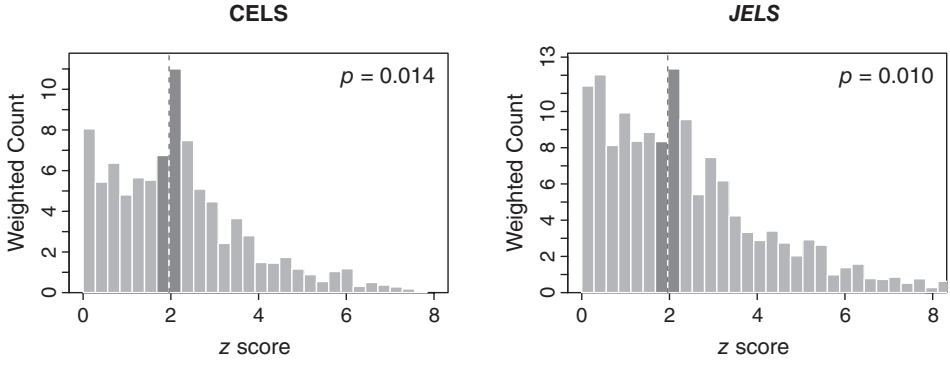
IV. Does Publication Bias Affect Empirical Legal Studies?

We find considerable evidence of publication bias. Figure 2 plots the distribution of z scores. In the absence of publication bias, we should expect the distribution of z scores to

[8]Note that our classification of observational studies is somewhat more expansive than conventional notions that would exclude descriptive studies with primary associations (but not causal inferences) of interest (see Rosenbaum 2002). Due to the small number of papers, we pooled these two categories, which could in principle be distinguished.

[9]Results essentially the same for the subset of studies presenting only t statistics of regression coefficients (the largest category of test statistics).

Figure 2: *z* scores from CELS 2012 and *JELS* 2004–2013.



NOTE: To present the distribution across papers, each *z* score is weighted inversely to the number of statistics from the respective underlying paper. We apply a similar caliper test as in Gerber and Malhotra (2008b), using the binomial distribution as a reference distribution to test for the difference in the weighted number of results just above and below the critical *z* score of 1.96 (plotted by the vertical line, with bins that are included in the caliper test denoted by darker shade). To weight papers equally when the number of hypothesis tests differs across papers, we calculate the one-tailed *p* value using a weighted binomial test by 100,000 Monte Carlo simulations. The *x*-axes are truncated at 8 for visibility.

be smooth around the threshold of 1.96. Instead, we find sharp discontinuities for CELS papers, with far more reported *z* scores just above 1.96 than just under. To formally test whether the discontinuity is statistically significant, we extend the "caliper test" proposed by Gerber and Malhotra (2008b) to weight papers equally.[10] The caliper is represented by the bins immediately above and immediately below the 1.96 threshold—the dark gray bins adjacent to the 1.96 threshold indicated by the vertical line. Under the null hypothesis, the probability of observing the discontinuity around the threshold is low for CELS papers (p value = 0.01). Interestingly, the relative density of *z* scores below 1.96 is much higher for the journal,[11] although the discontinuity at the threshold still points to publication bias (p value = 0.01). Conference bias is, if anything, more pronounced than publication bias.

---

[10]Gerber and Malhotra propose a one-tailed binomial test, weighing *z* scores equally. Weighting is particularly important in our extension because descriptive papers often yield many test statistics per paper, and we do not want a few papers to disproportionately drive the findings. To weight papers equally, we thereby weight *z* scores inversely to the number of *z* scores per paper, and take the difference between the number of weighted *z* scores above and below the threshold (within a 0.28 caliper, the bin size of the histogram in Figure 2). We then calculate the null distribution via 100,000 Monte Carlo simulations. The simulated (one-tailed) p value is the proportion of times that the simulated weighted difference is greater than or equal to the observed weighted difference (of *z* scores above and below the threshold). One challenge to the caliper test is that it is not obvious what the precise reference distribution in the absence of publication bias should be. Consider a reference distribution of a standard normal distribution, left truncated at the origin. By construction, the density is higher just below the critical value of 1.96 than above it. In that scenario, the one-tailed test may be conservative, as we would expect more *z* scores below the threshold. As Gerber and Malhotra (2008b) note, however, absent publication bias, we should expect the distribution of test statistics to be smooth around the threshold. Another challenge is that the test may overreject in instances where coefficient estimates are highly correlated within the same paper. Where possible, we collected *z* scores on the primary inference of interest to reduce this intrapaper correlation.

[11]A Kolmogorov-Smirnov test rejects the null of distributional equality (p value = 0.01).

Table 1: Breakdown of Types of Papers at CELS and *JELS*

	CELS		JELS	
	No.	*Prop.*	*No.*	*Prop.*
Outside of inclusion criteria	22	0.17	93	0.36
Descriptive (no primary inference)	55	0.20	64	0.25
Observational (primary inference)	53	0.42	73	0.28
Experimental	26	0.21	27	0.11

NOTE: Papers outside of inclusion criteria typically did not report test statistics. "No." indicates the number and "Prop." indicates the proportion.

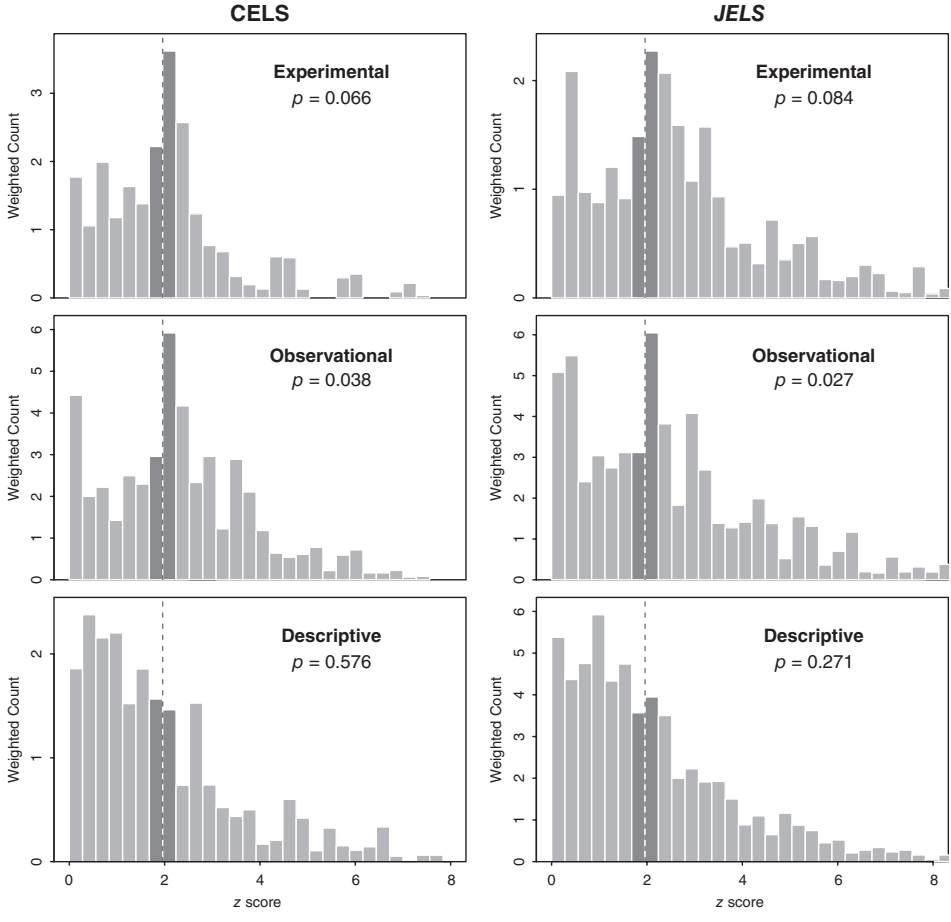
Why should this be the case? We consider several explanations. First, although CELS and *JELS* are closely related (the papers published in this issue after all were all presented at CELS), some conference papers might actually be quite distinct from papers submitted to *JELS*. For instance, the conference draws heavily on disciplinary papers in finance and political science. The incentive for authors from these disciplines to submit to *JELS* may be lower; and publication standards (and the primacy of hypothesis testing) may differ across disciplines. *JELS* also publishes relatively more descriptive work than was presented at the conference, and such work—by focusing less on specific parameters—may be less susceptible to publication bias. Table 1 breaks out our paper classifications by CELS in the left columns and *JELS* in the right columns: 36 percent of *JELS* papers fall outside our inclusion criteria, not engaging in formal statistical tests.[12] Relative to CELS, fewer *JELS* papers focus on primary inferences (e.g., causal effects about particular interventions) of interest. We can reject the hypothesis that paper types are identical across CELS and *JELS* (χ^2 test p value < 0.001).

Figure 3 confirms that publication bias differs across types of work. Each panel plots the distribution of z scores by type of paper. Experiments and observational studies exhibit discontinuities for both the journal and conference. The bottom panels present distributions for descriptive studies: in contrast to the top two rows, far more test statistics fall below the threshold and there is no statistically distinguishable discontinuity around the threshold. In short, *JELS*'s greater emphasis on descriptive work explains much of the difference in Figure 2.

A second explanation for the divergence between the conference and journal may lie in the conference review process: perhaps conference reviewers relied on a cut-off rule, while journal reviewers did not. This asymmetry, however, strikes us as unlikely. *JELS* draws on many of the same papers as well as reviewers. Moreover, the conference review process focused principally on research design and whether the paper, broadly speaking, fit under the rubric of empirical legal studies (e.g., no pure theory papers, no papers with attenuated connections to law). Of the reviews for some 360 papers, only 20 mentioned statistical significance; and even in those decisions, reviewer recommendations were nearly always based on research design. Indeed, while we attempted to collect the same information

[12]A small number of papers were also excluded because they did not present sufficient information to convert statistics to z scores.

Figure 3: z scores by type of paper.



NOTE: The left panels present statistics from CELS 2012 and the right panels present statistics from *JELS* 2004–2013. *P* values are calculated by a weighted binomial test from 100,000 Monte Carlo simulations. The dashed vertical line indicates the conventional critical z score of 1.96 and values falling in the dark bins are included in the caliper test. Axes are truncated at 8 for visibility. In the bottom right panel, the p value is roughly 0.27, despite the fact that the weighted counts are comparable, because the effective sample size within the caliper (taking into account weights) is small.

about test statistics for rejected papers, this became infeasible because many rejected papers did not provide enough information to be able to quantify statistical uncertainty or state a specific research question.

The last explanation paints the conference and journal review processes in a more salutary light. Much of the intellectual role of CELS and the refereeing process of *JELS* is to provide authors with feedback. If such feedback pushes back against specifications that happen to meet statistical thresholds, conference papers barely meeting statistical significance levels may appear quite differently in *JELS*. To examine this possibility, we compared versions of papers presented at CELS (whenever available) with versions published in *JELS*

conference issues.[13] Although the sample is very small, we find some evidence of this salutary role. For some papers, authors report *smaller* test statistics in the journal by applying more appropriate (and conservative) tests, such as clustering standard errors with panel data. In other instances, papers that were just above conventional thresholds at CELS improved the precision of estimates published in *JELS* by collecting more data. Tracking papers from conference to publication suggests that naïve notions of editors deploying cut-off rules are wrong. The evidence is more consistent with "satisficing" behavior by authors, which appears particularly acute for conferences.[14]

Lastly, it is also possible that *JELS* is simply unique in its editorial policy. As one co-editor of the journal, in exemplary fashion, noted: "*JELS* does not reject articles simply because of insignificant results."[15]

V. Conclusion

As in other fields, publication bias poses a considerable challenge to empirical legal studies. Drawing on both conference and journal materials suggests that the practice appears to be more endemic than a simple journal cut-off rule. Bias is, if anything, worse at conferences. In contrast to other disciplines, publication bias appears mitigated to some extent in empirical legal studies by considerable (and valuable) descriptive research. Yet even if *JELS* appears less susceptible than journals from other cognate disciplines, evidence of bias persists in experiments and observational studies. Our findings also strongly suggest that drawing on conference proceedings (e.g., in literature reviews, meta-analyses) will not cure publication bias.

While many other solutions have been proposed, one step in the right direction, echoing the same *JELS* co-editor and many others (e.g., Gerber & Malhotra 2008a), is simple: reviewers should focus on assessing the credibility of a study's research design and authors should present substantive effects with measures of statistical uncertainty—all regardless of whether results achieve conventional threshold levels of statistical significance. After all, none other than Ronald Fisher, founding father of hypothesis testing, called the calculation with fixed levels of significance "absurdly academic" (Fisher 1956).[16]

References

Blumenthal, Jeremy A. (2007) "Meta-Analysis: A Primer for Legal Scholars," 80 *Temple Law Rev.* 201.

---

[13]Due to the publication schedule of this issue, we were not able to do so for all papers published here.

[14]See McCrary (2006).

[15]Comment by Theodore Eisenberg, Empirical Legal Studies Blog, Nov. 21, 2008. As Eisenberg also noted privately, papers in which a key finding is the absence of statistical significance would also be well advised to assess Type II error.

[16]Nor were Neyman and Pearson fond of fixed significance levels (see Lehmann 1993).

Brodeur, Abel, Mathias Le, Marc Sangnier, & Yanos Zylberberg (2012) "Star Wars: The Empirics Strike Back," *Paris School of Economics Working Paper* 2012-29, available at <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2089580>.

Callaham M. L., & R. R. Wears (1998) "Positive-Outcome Bias and Other Limitations in the Outcome of Research Abstracts Submitted to a Scientific Meeting," 280(3) *JAMA* 254. doi:10.1001/jama.280.3.254.

Card, David, & Alan B. Krueger (1995) "Time-Series Minimum-Wage Studies: A Meta-Analysis," 85(2) *American Economic Rev.* 238. doi:10.2307/2117925.

Cochran, William G. (1952) "The χ^2 Test of Goodness of Fit," 23(3) *Annals of Mathematical Statistics* 315.

Donohue, John J., & Justin J. Wolfers (2006) "Uses and Abuses of Empirical Evidence in the Death Penalty Debate," 58(1) *Stanford Law Rev.* 791.

Easterbrook, P. J., R. Gopalan, J. A. Berlin, & D. R. Matthews (1991) "Publication Bias in Clinical Research," 337(8746) *Lancet* 867.

Fisher, Ronald A. (1956) *Statistical Methods and Scientific Inference.* Edinburgh: Oliver & Boyd.

Gerber, Alan S., & Neil Malhotra (2008a) "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals," 3(3) *Q. J. of Political Science* 313.

—— (2008b) "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" 37(1) *Sociological Methods & Research* 3. doi:10.1177/0049124108318973.

Gigerenzer, Gerd (2004) "Mindless Statistics," 33(5) *J. of Socio-Economics* 587.

Glass, Gene V., Barry McGaw, & Mary Lee Smith (1981) *Meta-Analysis in Social Research*, vol. 56. Beverley Hills, CA: Sage Publications.

Hopewell, Sally, Mike Clarke, & Sue Mallett (2005) "Grey Literature and Systematic Reviews," in H. R. Rothtstein, A. J. Sutton, & M. Borenstein, eds., *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, pp. 49–72. Chichester, West Sussex, UK: Wiley.

Leamer, Edward E. (1983) "Let's Take the Con Out of Econometrics," 73(1) *American Economic Rev.* 31.

Lehmann, Erich L. (1993) "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" 88(424) *J. of the American Statistical Association* 1242.

Masicampo, E. J., & Daniel R. Lalande (2012) "A Peculiar Prevalence of p Values Just Below .05," 65(11) *Q. J. of Experimental Psychology* 2271. doi:10.1080/17470218.2012.711335.

McCrary, Justin (2006) "Conservative Tests Under Satisficing Models of Publication Bias," available at <http://emlab.berkeley.edu/~jmccrary/mccrary2006_full.pdf>.

Patefield, W. M. (1981) "Algorithm AS 159: An Efficient Method of Generating Random R× C Tables with Given Row and Column Totals," 30(1) *J. of the Royal Statistical Society. Series C (Applied Statistics)* 91.

Petticrew, Mark, Simon Gilbody, & Fujian Song (1999) "Lost Information? The Fate of Papers Presented at the 40th Society for Social Medicine Conference," 53(7) *J. of Epidemiology & Community Health* 442.

Pfaff, John (2010) "A Plea for More Aggregation: The Looming Threat to Empirical Legal Scholarship," available at SSRN 1641435.

Rosenbaum, Paul R. (2002) *Observational Studies.* New York: Springer.

Rosenthal, Robert (1979) "The File Drawer Problem and Tolerance for Null Results," 86(3) *Psychological Bulletin* 638.

Roth, Alvin E. (1994) "Lets Keep the Con Out of Experimental Econ.: A Methodological Note," 19(2) *Empirical Economics* 279. doi:10.1007/BF01175875.

Song, A. Eastwood (1999) "The Role of Electronic Journals in Reducing Publication Bias," 24(3) *Informatics for Health & Social Care* 223.

Sporer, Siegfried, & Jane Goodman-Delahunty (2011) "Publication Bias in Meta-Analyses in Psychology and Law," presented as a conference paper at the Annual Meeting of the American Psychology-Law Society during the 4th International Conference on Psychology and Law.

Sterling, Theodore D. (1959) "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa," 54(285) *J. of the American Statistical Association* 30. doi:10.2307/2282137.