COMMENTARIES

Commentaries are informative essays dealing with viewpoints of statistical practice, statistical education, and other topics considered to be of general interest to the broad readership of *The American Statistician*. Commentaries are similar in spirit to Letters to the Editor, but they involve longer discussions of background, issues, and perspective. All commentaries will be refereed for their merit and compatibility with these criteria.

Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa

T. D. STERLING, W. L. ROSENBAUM and J. J. WEINKAM

This article presents evidence that published results of scientific investigations are not a representative sample of results of all scientific studies. Research studies from 11 major journals demonstrate the existence of biases that favor studies that observe effects that, on statistical evaluation, have a low probability of erroneously rejecting the so-called null hypothesis (H_0). This practice makes the probability of erroneously rejecting H_0 different for the reader than for the investigator. It introduces two biases in the interpretation of the scientific literature: one due to multiple repetition of studies with false hypothesis, and one due to failure to publish smaller and less significant outcomes of tests of a true hypotheses. These practices distort the results of literature surveys and of meta-analyses. These results also indicate that practice leading to publication bias have not changed over a period of 30 years.

KEY WORDS: Bias; Null results; Publication bias; Tests of significance.

1. INTRODUCTION

Scientific investigators are faced with the problem of interpreting consistent or contradictory findings of independent studies designed to test the same scientific hypothesis but usually conducted under different conditions. [Throughout this discussion we use the term scientific hypothesis to refer to the test of questions concerning the effect or lack of effect of a set of antecedent conditions to distinguish it from the corresponding statistical test of a null hypothesis (Kruskal and Tanur 1978).] Clinical investigators seek to determine from similar clinical trials if results of repeated applications of a therapeutic agent differ from chance. More recently, meta-analytic methods for summarizing the results of several studies have been developed that proceed under the (often implicit) assumption that all (or at least a representative sample) of relevant studies are available for analysis. Because access to study results is typically limited to published studies, the question of whether or not published studies constitute a representative sample of relevant studies is of concern.

The suggestion that published results of scientific investigations are not a representative sample of results from all scientific studies was first made in 1959 (Sterling 1959). Sterling found that 97.3% of papers published in four major psychology journals reported statistically significant outcomes for their major scientific hypotheses. Similar results (94%) were obtained by Bozarth and Roberts (1972). The possibility that results of published clinical trials may not be representative of all observed results was demonstrated by Simes (1986a; 1986b). They found that when only published trials were considered, the use of combined chemotherapeutic regimens for ovarian cancer was statistically significantly superior to the use of a single alkylating agent. However, when all registered trials (published and unpublished) were considered, the statistically significant advantage of combining chemotherapies disappeared. Dickersin et al. reported that 55% of published trials, compared with 15% of unpublished trials, had statistically significant results favoring a new therapy (Dickersin 1990; Dickersin, Chan, Chalmers, Sacks, and Smith 1987). Begg and Berlin (1988) give an example in which a positive result published in a prestigious journal continued to influence clinical practice even after the results were shown to be unreliable in publications that appeared subsequently in less prestigious journals. In an investigation of 11 meta-analyses of psychological topics, Glass, McGaw, and Smith (1981) found the average experimental effect from studies published in journals to be larger than the corresponding effect estimated from theses and dissertations. Coursol and Wagner (1986) found that researchers were less likely to submit—and journal editors less likely to publish studies—that were negative or neutral in outcome, resulting in different proportions of positive outcomes among published as compared to unpublished sources. Mahoney (1977) found that reviewers were highly influenced by the direction and strength of a submitted study's results. Sommer (1987) found that among studies conducted by members of the Society for Menstrual Research, proportionately more published studies reported statistically significant results than did unpublished studies. (Similar results have been reported by Dickersin 1992; Dickersin and Min 1993; Easterbrook, Berlin, Gopalan, and Matthews 1991.) Berlin, Begg, and Louis (1989) reported a substantial tendency for studies

T. D. Sterling and J. J. Weinkam are Professors, and W. L. Rosenbaum is Senior Research Associate, School of Computing Science, Faculty of Applied Sciences, Simon Fraser University, Burnaby, B. C., Canada V5A 1S6. The authors thank K. L. Jang for meticulously screening the research reports contained in numerous scientific journals. They are beholden also to Ingram Olkin for encouragement to resubmit this manuscript for publication.

with small sample sizes to report greater treatment effects than studies with larger samples.

These observations illustrate how the criteria used in the selection of papers for publication may operate as a filter preferring positive to negative results and thereby introduce a bias in their interpretation by readers. This may be because investigators may decide not to submit negative results or because journal referees and editors may be more likely to reject a submitted article if its results are negative.

(We cannot resist including a telling example. A letter from an editor of a major environmental/toxicological journal rejecting a manuscript because of its negative findings included the statement:

> Unfortunately, we are not able to publish this manuscript. The manuscript is very well written and the study was well documented. Unfortunately, the negative results translates into a minimal contribution to the field. We encourage you to continue your work in this area and we will be glad to consider additional manuscripts that you may prepare in the future.

A copy of this letter was submitted in confidence to the editor of *The American Statistician*.)

Moreover, publication bias makes the probabilities of Type I and Type II errors for a single study different for the reader than for the original investigator (Denton 1990). This bias is exacerbated when meta-analytic techniques are applied only to published studies because the analyses are likely to produce biased summary estimates that are apparently precise and accurate, leading to conclusions that may not only be wrong but at the same time may appear convincing (Begg and Berlin 1988). Furthermore, the publication criteria operate to include more than the true proportion of studies with nonnull results and less than the true proportion with null results. Such serious consequences have led to considerable concern. Chalmers (1991) has gone so far as to consider the system that results in publication bias a form of scientific misconduct.

In this report we update and expand Sterling's 1959 study in order to address the following questions: First, are present publication patterns consistent with publication bias? Second, have there been any changes in publication bias since its recognition about 30 years ago?

2. METHOD

The same four fields in psychology were reviewed as in 1958. However, the field of experimental psychology is now covered by the journals *General Psychology*; *Learning, Memory and Cognition*; *Human Perception and Performance*; and *Animal Behavior Processes*. The field of comparative and physiological psychology is now covered by the journals *Behavioral Neuroscience and Comparative Psychology*. The fields of clinical and social psychology are still covered by *The Journal of Consulting and Clinical Psychology* and the *Journal of Personality and Social Psychology*, respectively. All studies reviewed were published in 1987, except for those in *Animal Behavior Processes*, which were published in 1986.

The 1986 volumes of three medical journals were also reviewed. These were *The American Journal of Epidemiology*, *The American Journal of Public Health*, and *The New England Journal of Medicine*.

Only research papers that investigated a scientific hypothesis and actively collected data were included in this review. The criteria used for classifying the studies were the same as those used 35 years ago and were described in Sterling (1959). All eligible papers published during a calendar year were classified by a single reviewer according to whether or not the author of the published study decided that its major antecedent variables had resulted in an effect that was unlikely to be a chance result. What the author judged to be the major hypotheses and results were extracted from the paper's abstract or summary without taking strength of statistical significance into account. In the case of multiple tests the reviewer used his best judgment to select that test that appeared to be crucial to the final conclusion.

3. RESULTS

Table 1 summarizes the outcomes of tests of significance for the four fields of psychology (eight journals) and the three medical research journals. The table shows the number of research articles reviewed in 1986–1987; the percent of articles reviewed that used statistical tests of significance; the percent of articles using tests of significance

Table 1. Outcomes of Tests of Significance for Four Psychology and Three Medical Research Journals

Journals	No. of articles reviewed in 1986–87	% articles reviewed that use tests in 1986–87	% articles using tests that reject H_0 in 1986–1987	No. of articles reviewed that used tests in 1958	% articles using tests that reject H_0 in 1958
Experimental Psychology (four journals)	165	92.73	93.46	106	99.06
Comparative & Physiological Psychology (two journals)	119	88.24	97.14	94	96.81
Consulting & Clinical Psychology	83	96.39	97.50	62	95.16
Personality & Social Psychology	230	97.83	95.56	32	96.88
Psychology Journals Total	**597**	**94.30**	**95.56**	**294**	**97.28**
American Journal of Epidemiology	141	81.56	80.87	N/A	N/A
American Journal of Public Health	97	43.30	88.10	N/A	N/A
New England Journal of Medicine	218	75.69	87.88	N/A	N/A
Medical Journals Total	**456**	**69.25**	**85.40**	N/A	**N/A**

that rejected H_0 for their major hypotheses with $\alpha \leq .05$, and for comparison, the percent of articles using statistical tests that rejected H_0 in the 1958 study. For example, of the 165 articles reviewed in the 4 experimental psychology journals, 92.7% used tests of significance. Of those using tests of significance, 93.5% rejected H_0 for their principal hypothesis with $\alpha \leq .05$. The same decision was made in 1958 by 99.1% of research reports that used a test of significance. Overall, the publication practices of psychology journals appear to have remained unchanged since 1958 (with the possible exception of *The Journal of Experimental Psychology*). There were some differences between psychological and medical journal practices. In general, 94.3% of articles in the psychology journals used tests of significance, but only 69.2% of articles in the medical journals did. Of the medical journals, the investigators publishing in the *American Journal of Public Health* made least use of tests of significance. However, if medical investigators used tests of significance, H_0 was usually rejected, although not as often as in psychology journals.

It is interesting that we did not find any authors who expressly pointed out that they had rejected H_0 for their major hypotheses when obtained significance was larger than but still close enough to the conventional .05 (e.g., .07 or .08). There is also evidence that the use of the 5% significance level as a decision criterion is quite common among social scientists in general (McNemar 1960; Rosenthal 1979; Rosenthal and Gaito 1963; Smart 1964). It appears from sampling scientific journals in different subjects, especially in astronomy and chemistry, that statistical hypothesis testing is used in these sciences and that the practice of using $\alpha = .05$ as the dividing line between effect and no effect is practically universal.

4. DISCUSSION

The true situation is that scientific papers usually report the outcome of a number of scientific observations and include multiple statistical tests. Offhand it seems impossible to select objectively the major hypothesis and test of significance. Our analysis is based on having chosen from each published paper a single test of significance of the major hypothesis. While our data therefore are somewhat subjective, we feel that the strength of the results warrants the conclusions.

Suppose that several investigators have all performed experiments. Let α be the significance level of the tests, for example, the probability of Type I error, usually set at .05. Let β be the probability of Type II error. Because β in general is different for each experiment and depends on sample size and other factors, we shall use B as the average probability of Type II error for the experiments in question. Finally, let γ be the proportion of all scientific hypotheses tested for which the null hypothesis is really false (the collective batting average of the investigators).

The expected proportion of studies accepting H_0 will be $\gamma B + (1 - \gamma)(1 - \alpha)$.

If the set of published studies is a representative sample of the set of all conducted studies, then the proportion of published studies rejecting H_0 will be the same as the proportion of all conducted studies rejecting H_0.

Now α is usually .05 or less, and β, while unknown and variable, is frequently .15–.75 (Hedges 1984). For example, if $\alpha = .05$ and we take $B = .2$ as a conservative estimate, then the proportion of studies that should accept H_0 is $.95–.75\gamma$. Thus even if $\gamma = 1$, we would expect about 20% of published studies to be unable to reject H_0. Certainly $\gamma < 1$. In complex cases in which there are competing theories, there may well be as many (or more) false hypotheses as true ones.

In fact, the true value of γ is considerably smaller than 1 and the true value of B is undoubtedly considerably larger than .2. Consequently, the true proportion of experiments that should accept H_0 is considerably larger than .2.

The fact that in the literature the proportion of studies accepting H_0 is in general less than .2 casts serious doubt on the representativeness of the published studies as a sample of all studies. If we take this formula at face value, it suggests that only studies with high power are performed and that the investigators formulate only true hypothesis. Common experience tells us that such is unlikely.

From the perspective of a consumer of the scientific literature, especially for practitioners of meta-analysis, there are two basically different situations depending on whether or not the scientific hypothesis is true: the false hypothesis bias and the true hypothesis bias.

4.1 The False Hypothesis Bias

Because negative results are seldom published, either because they are not submitted for publication or because they are rejected in the reviewing process, there is nothing to prevent numerous replications of an experiment that in the long run should yield a negative result. Eventually, one of these replications will yield a significant result by chance, which will then have a higher probability of being published than replications terminating in negative results. Some safeguards exist to lessen the effect of false hypothesis bias. The publication of such results might elicit comments by others who had negative results in prior replications. Such comments however, usually appear in print long after the original article has been published. Another, probably more effective, safeguard is that investigators interested in the same sets of hypotheses might know each other and thus be able to compare differing results. These factors might decrease the possibility of false hypothesis bias remaining undetected but would not eliminate it entirely. In some cases of repeated clinical trials, registers are used to centralize information about replication to eliminate the effect of false hypothesis bias (Begg and Berlin 1988, 1989; Chalmers et al. 1986; Easterbrook 1987; Simes 1986a, 1986b). If enough replications are performed, some of the false positives will yield results greater than the null value, and these results may be published. Thus the point estimate produced by a meta-analysis will be biased and its confidence interval will be lengthened (unless the investigators were all influenced by what we might call a preconceived notion bias to perform one-sided tests in the same direction).

4.2 A True Hypothesis Bias

On the other hand, if the scientific hypothesis is true, most experiments testing that hypothesis will yield posi-

tive results, and many of them may be published. Some experiments, however, will yield negative or weak results, and will be relatively less likely to appear in print. Thus the point estimate obtained from a meta-analysis will be biased away from the null (Denton 1990), and its confidence interval will be shortened.

In practice, there are no firm standards regarding the quality necessary for a study to be included in the pool being meta-analyzed, or how different experimental or observational conditions can be before studies are rendered ineligible for inclusion. A number of techniques have been devised to overcome the serious limitation of any meta-analysis based solely on published or disseminated literature (Hetherington et al. 1989; Light and Pillemer, 1984; Smart 1964). None of these available methods is entirely satisfactory for dealing with publication bias. Meta-analysis for clinical trials may still be applicable if there is some assurance that all clinical trials are in a registry, and thus their outcomes may be comparable regardless of whether or not they have been submitted for publication or a report has been written for their dissemination (Dickersin and Berlin 1992; Dickersin, Min, and Meinert 1992).

5. BLIND-TO-OUTCOME PEER REVIEW

The influence of the outcome of a statistical test on the decision to publish scientific results is unsatisfactory for the following three reasons:

(1) It creates an uncertainty of how to interpret the outcome of a statistical test.

(2) It creates a misplaced impression of the relationship of statistical tests to scientific importance. (In a 1989 review of what is deemed important in scientific literature, Kruskal and Majors conclude with "we were depressed by the frequency of use of statistical significance as a measure of relative importance.")

(3) Present practices fail to inform on true null results. There are many instances where it is just as important to know the experimental conditions that do not produce effects as it is to know those that do.

A number of discussions of publication bias involve the determination of its causes (Begg and Berlin 1988; Dickersin and Berlin 1992; Sharp 1990), reducing its effects (Chalmers, Frank, and Reitman 1990; Easterbrook 1987; Newcombe 1987; Simes 1986b; Sharp 1990), and correcting meta-analyses for its effects (Begg and Berlin 1988; Denton 1990; Hedges 1984; Hedges and Olkin 1985; Hunter 1990; Iyengar and Greenhouse 1988). Although a thorough discussion of remedies for publication bias is beyond the scope of this paper, one possible method of greatly diminishing the influence of publication bias is to accept or reject a scientific study without paying attention to, or perhaps even in ignorance of, its outcome a blind-to outcome peer review. In this strategy, editors or reviewers base their decision on:

(1) the importance of the study (justified in the introductory section by a description of planned work, its history, importance, and hypothesis to be tested, etc.); and

(2) the relevance of the proposed methods and of the data to be obtained for the purpose of the scientific enquiry.

This reviewing process could be done either before or after the actual study. Several related suggestions have been made (Begg 1985; Kochor 1986; Kupfersmid 1988; Kupfersmid, personal communication, 1990; Newcombe 1987; Rosenthal, 1966). The merits of such radical changes in editorial policy have been discussed in detail by Begg and Berlin (1989). In short, more radical measures than public consciousness raising are needed to curtail the influence of publication bias.

[*Received August 1993. Revised April 1994.*]

REFERENCES

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

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

——— (1989), "Publication Bias and Dissemination of Clinical Research," *Journal of the National Cancer Institute*, 81, 2, 107–115.

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

Bozarth, J. D., and Roberts, R. R. (1972), "Signifying Significant Significance," *American Psychologist*, 27, 774–775.

Chalmers, I. (1991), "Under-Reporting Research is Scientific Misconduct," *Journal of the American Medical Association*, 2263, 1405–1408.

Chalmers, I., et al. (1986), "The Oxford Database of Perinatal Trials: Developing a Register of Published Reports of Controlled Trials," *Controlled Clinical Trials*, 7, 306–324.

Chalmers, T. C., Frank, C. S., and Reitman, D. (1990), "Minimizing the Three Stages of Publication Bias," *Journal of the American Medical Association*, 263, 1392–1395.

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

Denton, F. T. (1990), "The Effects of Publication Selection Test Probabilities and Estimator Distributions," *Risk Analysis*, 10, 131–136.

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

———(1992), "Why Register Clinical Trials?—Revisited," *Controlled Clinical Trials*, 13, 170–177.

Dickersin, K., and Berlin, J. A. (1992), "Meta-Analysis: State-of-the-Science," *Epidemiology Review*, 14, 154–176.

Dickersin, K., and Min, Y. I. (1993), "Publication Bias: The Problem That Won't Go Away," *Annals of the New York Academy of Science*.

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., and Smith, H. (1987), "Publication Bias and Clinical Trials," *Controlled Clinical Trials*, 8, 343–353.

Dickersin, K., Min, Y. I., and Meinert, C. L. (1992), "Factors Influencing Publication of Research Results: Follow-Up of Applications Submitted to Two Institutional Review Boards," *Journal of the American Medical Association*, 267, 374–378.

Easterbrook, P. (1987), "Reducing Publication Bias," *British Medical Journal*, 295, 1347.

Easterbrook, P., Berlin, J. A., Gopalan, R., and Matthews, D. R. (1991), "Publication Bias in Clinical Research," *The Lancet*, 337, 867–872.

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

Hedges, L. 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.

Hedges, L. V., and Olkin, I. (1985), *Statistical Methods for Meta-Analysis*, Orlando, FL: Academic Press.

Hetherington, J., et al. (1989), "Retrospective and Prospective Identification of Unpublished Controlled Trials: Lessons from a Survey of Obstetricians and Pediatricians," *Pediatrics*, 84, 374–380.

Hunter, J. E. (1990), *Methods of Meta-Analysis: Correcting Bias and Error in Research Findings*, Newbury Park, CA: Sage Publications.
Iyengar, S., and Greenhouse, J. B. (1988), "Selection Models and the File Drawer Problem," *Statistical Science*, 3, 9–135.
Kochor, M. S. (1986), "The Peer Review of Manuscripts in Need of Improvement," *Journals of Chronic Diseases*, 39, 147–149.
Kruskal, W. H., and Majors, R. (1989), "Concepts of Relative Importance in Recent Scientific Literature," *The American Statistician*, 3, 1, 2.
Kruskal, W. H., and Tanur, J. M. (1978), "Significance, Tests Of," *International Encyclopedia of Statistics*, 2, 944–956.
Kupfersmid, J. (1988), "Improving What is Published: A Model in Search of an Editor," *American Psychologist*, 635–642.
Light, R. J., and Pillemer, D. B. (1984), *Summing Up: The Science of Reviewing Research*, Cambridge, MA: Harvard University Press.
Mahoney, M. J. (1977), "Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System," *Cognitive Therapy Research*, 1, 161–175.
McNemar, Q. (1960), "At Random: Sense and Nonsense," *American Psychologist*, 15, 295–300.
Newcombe, R. G. (1987), "Towards a Reduction in Publication Bias," *British Medical Journal*, 295, 6566–6569.
Rosenthal, R. (1979), "The File Drawer Problem and Tolerance for Null Results," *Psychology Bulletin*, 86, 638–641.
Rosenthal, R., and Gaito, J. (1963), "The Interpretation of Levels of Significance by Psychological Research," *Journal of Psychology*, 55, 323–338.
Rosenthal, R. (1966), *Experimenter Effects in Behavioral Research*, New York: Appleton-Century-Crofts, pp. 35–37.
Sharp, D. A. (1990), "What Can and Should be Done to Reduce Publication Bias," *Journal of the American Medical Association*, 263, 1390–1391.
Simes, R. J. (1986a), "Confronting Publication Bias: A Cohort Design for Meta-Analysis," *Statistics in Medicine*, 6, 11–29.
——— (1986b), "Publication Bias: The Case for an International Registry of Clinical Trials," *Journal Clinical of Oncology*, 4, 1529–1541.
Smart, R. G. (1964), "The Importance of Negative Results in Psychological Research," *Canadian Psychologist*, 5, 225–232.
Sommer, B. (1987), "The File Drawer Effect and Publication Rates in Menstrual Cycle Research," *Psychology of Women Quarterly*, 11, 233–242.
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.