Are Null Results Becoming an Endangered Species in Marketing?

Raymond Hubbard
College of Business and Administration, Drake University, Des Moines, IA 50311

J. Scott Armstrong[*]
The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

Marketing Letters 3:2, (1992): 127-136 1992 Kluwer Academic Publishers, Manufactured in the Netherlands.

Key words: File Drawer Problem, Null Results, Publication Bias, Statistical Power Analysis, Statistical Significance

Abstract

Editorial procedures in the social and biomedical sciences are said to promote studies that falsely reject the null hypothesis. This problem may also exist in major marketing journals. Of 692 papers using statistical significance tests sampled from the *Journal of Marketing, Journal of Marketing Research,* and *Journal of Consumer Research* between 1974 and 1989, only 7.8% failed to reject the null hypothesis. The percentage of null results declined by one-half from the 1970s to the 1980s. The *JM* and the *JMR* registered marked decreases. The small percentage of insignificant results could not be explained as being due to inadequate statistical power.

Various scholars have claimed that editorial policies in the social and medical sciences are biased against studies reporting null results, and thus encourage the proliferation of Type 1 errors (erroneous rejection of the null hypothesis). Greenwald (1975, p. 15) maintains that Type I publication errors are underestimated to the extent that they are: ". . . frightening, even calling into question the scientific basis for much published literature."

Our paper examines the publication frequency of null results in marketing. First, we discuss how editorial policies might foster an atmosphere receptive to Type I error proliferation. Second, we review the evidence on the publication of null results in the social and biomedical sciences. Third, we report on an empirical investigation of the publication frequency of null results in the marketing literature. Fourth, we examine power levels for statistically insignificant findings in marketing to see if they are underpowered and thus less deserving of publication. Finally, we provide suggestions to facilitate the publication of null results.

1. Editorial policies and Type I error proliferation

[*] *The authors thank Richard Bagozzi, Gilbert Churchill, James Engel, John Farley, Anthony Greenwald, Daniel Vetter, the editor, and two anonymous reviewers for comments on earlier versions of this paper. Any errors remaining are our responsibility.

Some researchers allege that Type I errors will increase if editors display a bias to publish studies whose results are statistically significant at the .01 or .05 levels. Bakan (1966, p. 427), in commenting upon a former editor (Arthur W. Melton) of the *Journal of Experimental Psychology*, observed that: “His clearly expressed opinion that non-significant results should not take up the space of the journals is shared by most editors of psychological journals.” Others, such as psychologists (Greenwald 1975; Rosnow and Rosenthal 1989),economists (Feige 1975), and statisticians (Salsburg 1985), share this belief about a publication bias against null results.

The behavior of authors and editors might be influenced by the statistical significance of a study's findings. Three possibilities, outlined below, exist for a researcher who obtains null results.

1.1 Null results are not submitted for publication

Null results are less likely to be submitted for publication than are non-null results. Greenwald (1975), using intentions data from 36 *Journal of Personality and Social Psychology (JPSP)* authors, estimated the probability of null results being submitted for publication to be about one-tenth that had they obtained significant results (.06 compared to .59 probability). Coursol and Wagner's (1986) examination of 609 returns from a survey of counseling psychologists found that while 82% of the articles reporting positive outcomes were submitted for publication, only 43% of those with neutral or negative findings were submitted.

1.2. Null results are unlikely to be published

If submitted, null results are less likely to be published than their non-null counterparts. For example, Kerr, Tolliver, and Petree (1977) surveyed the editors and advisory board members of nineteen leading management and social science journals in order to elicit common reasons for manuscript acceptance or rejection. They concluded that even when a manuscript was judged to be otherwise competent and of current interest to the field, statistically insignificant results would result in a substantially lower likelihood of acceptance. Similar results were obtained in a survey of manuscript reviewers for Canadian psychology journals (Rowney and Zenisek 1980). Finally, Atkinson, Furlong, and Wampold (1982) asked 101 consulting editors of two psychology journals to evaluate three versions of a manuscript that differed only with regard to the level of statistical significance reported. The statistically nonsignificant and almost significant versions were more than three times as likely to be rejected for publication than was the statistically significant one.

Many authors share the beliefs of editors and reviewers. For example, 61% of authors who had published empirical articles in various education and psychology journals in 1988 believed that only research yielding statistically significant findings would be published (Kupfersmid and Fiala 1991).

Coursol and Wagner (1986) found that of 106 statistically significant papers submitted for publication in counseling psychology, 80% were accepted, but only half of 28 studies with null results were accepted. Sommer's (1987) survey of members of the Society for Menstrual

Cycle Research reported a 73% publication rate for papers with statistically significant outcomes, and a corresponding rate of 54% for those with null results.

Four previous empirical studies, all from psychology, suggest the existence of a bias against the publication of reports failing to reject the null hypothesis. For instance, Sterling's (1959) investigation of papers published in four leading psychology journals during 1955 found that only 2.7% of those using statistical significance tests failed to reject the null hypothesis. Smart's (1964) analysis of these same four journals in 1962 revealed that 8.7% reported statistically insignificant results. Bozarth and Roberts (1972) discovered that only 6% of all articles using statistical tests in three prominent counseling psychology journals published between January 1967 and August 1970 were unable to reject the null hypothesis. Greenwald's (1975) estimate, based on a content analysis of a single annual (1972) issue of *JPSP*, was 12.1%.

1.3. Further research by author

A third option for the individual faced with nonsignificant findings is to persevere with the topic. Greenwald (1975) asked researchers if they were likely to conduct an exact or modified replication of their work following an initial full-scale test of their main hypothesis. If the initial result was statistically significant, the probability was .36; if an insignificant result was obtained, this probability was .62.

Persevering with a research topic might include the reworking of results prior to submission for publication, so that at least portions of them are statistically significant. Some researchers might engage in data-mining activities, thus promoting Type I errors (Feige 1975).

2. The file drawer problem

Editorial policies, real or perceived, might contribute to what Rosenthal (1979, p. 638) calls the file drawer problem: "... journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (e.g., p > .05) results." The implications of this problem may show up when a meta-analysis is undertaken. Because unpublished (file drawer) works are often poorly represented in meta-analyses, published effect size estimates are inflated. For example, 10 of the 12 education and psychology meta-analyses reviewed by Smith (1980) showed average effect sizes in published journal accounts to be 33% higher than those reported in theses and dissertations. Shaddish, Doherty, and Montgomery (1989) asked a random sample of 519 members of organizations involved with family and marital psychotherapy outcomes if they possessed file drawer studies on the issue. After analyzing the 375 replies, they estimated that there may be almost as many of these unpublished works as there are published studies and dissertations. They concluded that population effect sizes of published works are about 10% to 40% larger than those computed from unpublished research. Using a maximum likelihood modeling approach, Rust, Lehmann, and Farley (1990) examined whether publication bias, based on effect sizes rather than statistical significance levels, was present in two published meta-analyses and a proprietary data set. They concluded that there was some bias in published consumer behavior experiments, as, on average, effect sizes were inflated by about 5%; effect sizes in advertising carryover models appeared to be inflated by 10%.

Evidence from medical studies also points to a publication bias against null outcomes. Simes (1986), for example, showed that whereas the pooled results of published trials for a certain treatment of ovarian cancer revealed statistically significant benefits, the pooled results of registered trials (which included both published and unpublished studies) evaluating the same treatment did not. Similarly, 318 authors of published clinical trials were surveyed to see if they had been involved with any unpublished trials (Dickersin et al. 1987). Responses from 156 individuals yielded 271 unpublished and 1,041 published trials; while only 14% of the unpublished studies favored the test therapy, this figure was 55% for the published reports.

3. Publication frequency of null results in marketing

How frequently are null results published in marketing? The first author conducted a content analysis of 32 randomly selected issues of each of the *Journal of Marketing (JM)*, *Journal of Marketing Research (JMR)*, and *Journal of Consumer Research (JCR)*. This represents a 50% random sample of all issues of *JM, JMR*, and *JCR* published between 1974 and 1989. (*JCR* was first published in 1974, hence the point of departure for the present study.)

Determination as to whether research reports rejected or failed to reject the null hypothesis was accomplished by employing Sterling's (1959, footnote 2, pages 31-32) detailed classification scheme. In common with Sterling, nearly all of the empirical studies examined here were multivariate in character. Thus, whenever a dominant hypothesis was evident, its statistical evaluation was recorded. When a study used two or more variables and it was not obvious as to which was the most important, if H_o was not rejected for at least half of these variables, the article was classified as H_o not rejected and vice versa. A similar rationale was adopted for studies that reported more than one experiment.

Of the 1,148 papers in our sample, 60.3% used significance tests in their analyses. Of the latter, 7.8% failed to reject the null hypothesis. (Inspection of a ten percent random sample of these papers by a colleague, Daniel Vetter, supported these findings.) In comparison with similar studies in psychology, this percentage is higher than that obtained by Sterling (2.7%) and Bozarth and Roberts (6.0%), and is close to Smart's (8.7%) estimate.

The three journals contained similar percentages of research reports failing to reject the null hypothesis. For the *JM* this figure was 6.8%, with values of 8.6% and 7.5% for *JMR* and *JCR* respectively.

The percentage of papers using significance tests has been increasing over time. Table I shows that this figure rose from 50.4% in 1974-1979 to 68.0% in 1980-1989. The percentage of papers with insignificant findings, however, declined by one-half from the 1970s to the 1980s (from 11.4% to 5.7%). Substantial decreases were noted for the *JM* and the *JMR*, while *JCR* showed a slight increase.

We cannot be sure whether the reduction in the incidence of null results is attributable to publication bias. It may be that this decrease is due to a desire for fairness in reviewing. Significant results meet an objective criterion for accepting a paper. If the number of submissions increases relative to the space available, one might expect this to result in a decreasing proportion of papers with null results. In addition, researchers may have become more skilled at designing studies that avoid null outcomes, or are less willing to submit manuscripts that do not reject the null hypothesis than they were in the 1970s. What is clear, however, is that the null result became a rare breed in the 1980s than it was during a good part of the preceding decade.

Table 1
Changes over time in papers failing to reject the null hypothesis

	1974-1979			1980-1989		
Journals	Research Reports	% Reports using significance tests	% Reports failing to reject H_o[b]	Research Reports	% Reports using significance tests	% Reports failing to reject H_o[b]
JM	179	29.1	13.5	190	50.5	3.1
JMR	214	60.3	13.2	217	74.7	4.9
JCR[a]	111	65.8	6.8	237	75.9	7.8
Totals	504	50.4	11.4	644	68.0	5.7

[a] Calculations for JCR are based on Volumes I through 6 inclusively, and thus incorporate the March 1980 issue.

[b] Represents the percentage of those reports that use significance tests.

4. Statistical power of nonsignificant outcomes

The power of a statistical test (the probability of rejecting a false null hypothesis) depends upon the significance criterion, the effect size in the population, and the sample size (Cohen 1988; Sawyer and Ball 1981). Tests with inadequate statistical power are more likely to yield null results, and thus are less deserving of publication. On the other hand, statistically nonsignificant outcomes accompanied by high power are potential contributions to knowledge (Fagley 1985). It is therefore important to examine the power of the nonsignificant results.

Cohen (1988) provides power tables, with various alpha levels, experimental size effects, and sample sizes, for a number of commonly employed statistical methods. These tables were used to calculate the statistical power exhibited by the studies in our sample that reported nonsignificant outcomes. Standard procedures outlined by Cohen (1962; 1988) were also followed here. That is, all power assessments involved (1) nondirectional tests, (2) a = .05, (3) only major statistical tests, with manipulation checks and peripheral reliability estimates omitted, (4) conventional definitions of small, medium, and large effect sizes, and (5) the article as the unit of analysis.[1]

4.1. Results

Eleven of the 54 articles with nonsignificant results could not be power-analyzed, six because they employed techniques for which power tests are unavailable, and five because they provided insufficient information. The 43 remaining articles allowed power analyses of 410 statistical tests, an average of 9.5 tests per article. Of these 43 studies, 8 were from *JM,* 16 from *JMR,* and 19 from *JCR.*

Table 2 presents the frequency and cumulative percentage distributions of the mean power of the 43 articles to detect small, medium, and large effect sizes. Cohen (1988) recommends that the value .80 be used when no other basis exists for establishing a desired power level. Given this recommendation, the average power levels of these articles is high; the chances of detecting small, medium, and large effects per article are .35, .89, and .99, respectively. Corresponding values based on the 410 individual tests are .36, .87, and .98. These power figures are virtually identical to those obtained by Sawyer and Ball (1981) from their analysis of 23 empirical articles published in the four 1979 issues of *JMR* (small effect = .41, medium effect = .89, and large effect = .98).

Table 2
Power of published studies with statistically nonsignificant outcomes
(cumulative percentages based on 43 studies)

	Small effects		Medium effects		Large effects	
Power	Frequency	Cumulative percentage	Frequency	Cumulative percentage	Frequency	Cumulative percentage
.99-			14	100.0	32	100.0
.95-.98			5	67.4	8	25.6
.90-.94			8	55.8	2	7.0
.80-.89	2	100.0	6	37.2	1	2.3
.70-.79	3	95.3	5	23.2		
.60-.69	5	88.3	4	11.6		
.50-.59	4	76.7	1	2.3		
.40-.49	0	67.5				
.30-.39	2	67.5				
.20-.29	14	62.8				
.10-.19	13	30.2				

For the period 1974-1989, all three journals displayed similar power levels to detect small, medium, and large effects. For *JM* these values were .29, .86, and .99; for *JMR* they were .33, .92, and .99; while for *JCR* they were .40, .87, and .98.

On average across all three journals, the nonsignificant studies included in the present sample had a modest possibility of detecting even small effect sizes. Fourteen of the 43 articles (32.6%) had a 50-50 chance or better of doing so, while two managed to exceed the suggested level of .80. If medium effect sizes in the marketing literature are assumed, these articles had an almost 90% chance, on average, of distinguishing them. Ten of the studies (23.3%) failed to meet the .80 recommended power benchmark (but nine of these exhibited power in the .60 to .79 range); 33% met or exceeded the .99 power level; and none were below the .50 level. All 43 of the articles met or surpassed the 80% power value to discern large effect sizes. Indeed, 32 of the studies (74.4%) had a .99 chance or greater of rejecting a false null hypothesis (Table 2).

The average power levels associated with articles reporting statistically nonsignificant outcomes were consistent over time. For example, during 1974-1979, the ability to uncover small, medium, and large effects across all three journals was .29, .88, and .99. For the period 1980-1989, these figures were .42, .90, and .99.

4.2. Did high statistical power facilitate publication of null results?

It might be argued that one reason these 43 papers with null results were published is their generally high levels of statistical power, the implication being that unpublished studies with null outcomes are noticeably underpowered. It was not possible to assess directly the merits of this argument because, to the best of our knowledge, the power levels of published and unpublished marketing papers with null results have never been compared.

Indirect evidence, however, suggests that concerns about power played an inconsequential role in the publication decision. Authors of published studies attaining, or failing to attain, statistically significant findings do not appear to formally incorporate power considerations into their research designs (Cohen 1990). In our study, none of the 54 articles with insignificant results presented power calculations; five of these studies were published without the information necessary to compute statistical power levels. Sawyer and Ball (1981) surveyed authors of empirical articles published in five issues of *JMR* (November 1978 to November 1979) about how they determined sample size. They concluded that the explicit computation of statistical power is not common among researchers; only 4 of 28 respondents (14%) calculated power before data collection, and an additional two respondents did so afterwards. A content analysis of these same five *JMR* issues by the present authors revealed that none of the 59 articles employing statistical significance tests reported power calculations, and only 4 alluded to the issue of adequate sample sizes.

5. Conclusions

Researchers in various areas of the social and biomedical sciences have shown what they claim is a bias against publishing works that fail to reject the null hypothesis. Our study found similar conditions in major marketing journals. Furthermore, this problem seems to be getting more serious over time as the publication frequency of null results declined by one-half from the 1970s to the 1980s. Is it possible that null results are an endangered species?

Studies that fail to reject the null hypothesis, but demonstrate explicitly that they meet or exceed the .80 power level recommended by Cohen (1988), might be evaluated for publication on the basis of whether they yield valuable information. Null results can provide evidence of a trivial effect size in the population. Conversely, statistically significant results accompanied by high power, often gained with large samples, may have negligible effect sizes. Thus, null outcomes can be meaningful. The publication of null results also might help to prevent researchers from reinvestigating blind-alleys. Recognizing the potential contribution of nonsignificant results, editorials in the *Journal of Clinical Neuropsychology* (Rourke and Costa 1979) and the *New England Journal of Medicine* (Angell 1989) have indicated the willingness of these journals to publish well-designed papers with null findings. Marketing journals might benefit by adopting similar policies.

References

Angell, Marcia. (1989). "Negative Studies," *New England Journal of Medicine,* 321, 464 466.

Atkinson, Donald R., Michael J. Furlong, and Bruce E. Wampold. (1982). "Statistical Significance, Reviewer Evaluations, and the Scientific Process: Is There a (Statistically) Significant Relationship?" *Journal of Counseling Psychology,* 29, 189-194.

Bakan, David. (1966). "The Test of Significance in Psychological Research," *Psychological Bulletin,* 77, 423 437.

Bozarth, Jerold D., and Ralph R. Roberts, Jr. (1972). "Signifying Significant Significance," *American Psychologist,* 27, 774-775.

Cohen, Jacob. (1962). "The Statistical Power of Abnormal-Social Psychological Research: A Review," *Journal of Abnormal and Social Psychology,* 65, 145-153.

Cohen, Jacob. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cohen, Jacob. (1990). "Things I Have Learned (So Far)," *American Psychologist*, 45, 1304-1312.

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

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

Fagley, N. S. (1985). "Applied Statistical Power Analysis and the Interpretation of Nonsignificant Results by Research Consumers," *Journal of Counseling Psychology,* 32, 391-396.

Feige, Edgar L. (1975). "The Consequences of Journal Editorial Policies and a Suggestion for Revision," *Journal of Political Economy,* 83, 1291-1295.

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

Kerr, Steven, James Tolliver, and Doretta Petree. (1977). "Manuscript Characteristics Which Influence Acceptance for Management and Social Science Journals," *Academy of Management Journal*, 20, 132-141.

Kupfersmid, Joel, and Michael Fiala. (1991). “A Survey of Attitudes and Behaviors of Authors Who Publish in Psychology and Education Journals,” *American Psychologist*, 46, 249-250.

Rosenthal, Robert. (1979). “The ‘File Drawer Problem’ and Tolerance for Null Results,” *Psychological Bulletin*, 86, 638-641.

Rosnow, Ralph L., and Robert Rosenthal. (1989). “Statistical Procedures and the Justification of Knowledge in Psychological Science,” *American Psychologist*, 44, 1276-1284.

Rourke, Byron P., and Louis Costa. (1979). “Editorial Policy 11,” *Journal of Clinical Neuropsychology*, 1, 93-95.

Rowney, Julie A., and Thomas J. Zenisek. (1980). “Manuscript Characteristics Influencing Reviewers' Decisions,” *Canadian Psychology*, 21, 17-21.

Rust, Roland T., Donald R. Lehmann, and John U. Farley. (1990). “Estimating Publication Bias in Meta-Analysis,” *Journal of Marketing Research*, 27, 220-226.

Salsburg, David S. (1985). “The Religion of Statistics as Practiced in Medical Journals,” *American Statistician*, 39, 220-223.

Sawyer, Alan G., and A. Dwayne Ball. (1981). “Statistical Power and Effect Size in Marketing Research,” *Journal of Marketing Research*, 18, 275-290.

Shaddish, William R., Maria Doherty, and Linda M. Montgomery. (1989). “How Many Studies Are in the File Drawer? An Estimate from the Family/Marital Psychotherapy Literature,” *Clinical Psychology Review*, 9, 589-603.

Simes, Robert J. (1986). “Publication Bias: The Case for an International Registry of Clinical Trials,” *Journal of Clinical Oncology*, 4, 1529-1541.

Smart, Reginald G. (1964). “The Importance of Negative Results in Psychological Research,” *Canadian Psychologist*, 5, 225-232.

Smith, Mary L. (1980). “Publication Bias and Meta-Analysis,” *Evaluation in Education*, 4, 22-24.

Sommer, Barbara. (1987). “The File Drawer Effect and Publication Rates in Menstrual Cycle Research,” *Psychology of Women Quarterly*, l l, 233-241.

Sterling, Theodore D. (1959). “Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance – or Vice Versa,” *Journal of the American Statistical Association*, 54, 30-34.

[1] The statistical tests include the t test, the significance of Pearson's r, the significance of the difference between correlation coefficients, the sign test, the test for differences between proportions, chi-square tests, the F test in ANOVA, ANCOVA, and multiple regression, and power tests for MANOVA and MANCOVA. The reader can consult Cohen (1988) for the details involved in calculating power levels for small, medium, and large effects. Also, see Sawyer and Ball (1981, p. 276) for a summary table of these (excluding MANOVA and MANCOVA).