Psychological Bulletin
1979, Vol. 86, No. 3, 638–641

The "File Drawer Problem" and Tolerance for Null Results

Robert Rosenthal
Harvard University

For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results. Quantitative procedures for computing the tolerance for filed and future null results are reported and illustrated, and the implications are discussed.

Both behavioral researchers and statisticians have long suspected that the studies published in the behavioral sciences are a biased sample of the studies that are actually carried out (Bakan, 1967; McNemar, 1960; Smart, 1964; Sterling, 1959). The extreme view of this problem, the "file drawer problem," is that the 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.

In the past there was very little one could do to assess the net effect of studies, tucked away in file drawers, that did not make the magic .05 level (Rosenthal & Gaito, 1963, 1964). Now, however, although no definitive solution to the problem is available, one can establish reasonable boundaries on the problem and estimate the degree of damage to any research conclusion that could be done by the file drawer problem.

This advance in our ability to cope with the file drawer is an outgrowth of the increasing interest of behavioral scientists in summarizing bodies of research literature systematically and quantitatively, both with respect to significance levels (Rosenthal, 1969, 1976, 1978) and with respect to effect-size estimation (Hall, 1978; Rosenthal, 1969, 1976; Rosenthal & Rosnow, 1975; Smith & Glass, 1977; Glass, Note 1). One hopes that this interest in summarizing entire research domains will lead to an improvement in bookkeeping so that eventually all results will be recorded both with an estimate of effect size (e.g., r or d; Cohen, 1977) and with the level of significance obtained, or more practically, with the standard normal deviate (Z) that corresponds to the obtained p (Rosenthal, 1978).[1] Future appraisals of research domains of the type found in *Psychological Bulletin* should give estimates of overall effect sizes and significance levels; these estimates of overall significance can provide a basis for coping with the file drawer problem.

Tolerance for Future Null Results

Given any systematic quantitative review of the literature bearing on a particular hy-

Preparation of this article was supported in part by the National Science Foundation.

I would like to thank Judith A. Hall and Donald B. Rubin for their valuable improvements of an earlier version of this article.

Requests for reprints should be sent to Robert Rosenthal, Department of Psychology and Social Relations, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138.

[1] Standard normal deviates (Z) can be found by various methods, of which the following three are most often useful: (a) Obtain the exact p associated with the test statistic (e.g., t, F, or χ^2) and find the Z associated with that p in tables of the normal distribution; (b) if the effect size r or phi is given or can be computed, Z can be estimated by $r(N)^{\frac{1}{2}}$; (c) if the effect size d is given or can be computed, Z can be estimated by $[d^2/(d^2+4)]^{\frac{1}{2}}(N)^{\frac{1}{2}}$.

pothesis, for example, that psychotherapy is effective (Glass, Note 1), that women are more sensitive than men to nonverbal cues (Hall, 1978), or that one person's expectation for another person's behavior can come to serve as self-fulfilling prophecy (Rosenthal, 1969, 1976), it is easy to calculate an overall probability, based on all the independent studies available to the reviewer, that the effect in question is "real," that is, not a Type I error (Rosenthal, 1978). The fundamental idea in coping with the file drawer problem is simply to calculate the number of studies averaging null results that must be in the file drawers before the overall probability of a Type I error is brought to any desired level of significance, say, $p = .05$. This number of filed studies, or the tolerance for future null results, is then evaluated for whether such a tolerance level is small enough to threaten the overall conclusion drawn by the reviewer. If the overall level of significance of the research review will be brought down to the level of *just significant* by the addition of just a few more null results, the finding is not resistant to the file drawer threat.

Computation

Perhaps the simplest, most useful way of computing the overall p of a set of research studies is the method of adding Zs (Cochran, 1954; Mosteller & Bush, 1954; Rosenthal, 1978). This method requires only that one add the standard normal deviates of Zs associated with the ps obtained and divide by the square root of the number of studies being combined. The result is itself a Z that can be entered in a table to find the associated overall p:

$$Z_c = k\bar{Z}_k/\sqrt{k} = \sqrt{k}\bar{Z}_k, \qquad (1)$$

where Z_c is the new combined Z, k is the number of studies combined, and $\bar{Z}_k$ is the mean Z obtained for the k studies.

To find the number (X) of new, filed, or unretrieved studies averaging null results required to bring the new overall p to any desired level, say, just significant at $p = .05$ ($Z = 1.645$), one simply writes:

$$1.645 = k\bar{Z}_k/\sqrt{k + X}. \qquad (2)$$

Rearrangement shows, then, that

$$X = (k/2.706)[k(\bar{Z}_k)^2 - 2.706]. \qquad (3)$$

An alternative formula that may be more convenient when the sum of the Zs (ΣZ) is given rather than the mean Z is as follows: $X = [(\Sigma Z)^2 / 2.706] - k$. One method based on counting rather than adding Zs may be easier to compute and can be employed when exact p levels are not available; but it is probably less powerful. If X is the number of new studies required to bring the overall p to .50 (not to .05), s is the number of summarized studies significant at $p < .05$, and n is the number of summarized studies not significant at .05, then $X = 19s - n$. Another conservative alternative when exact p levels are not available is to set $Z = .00$ for any nonsignificant result and to set $Z = 1.645$ for any result significant at $p \leq .05$.

Equations 1, 2, and 3 all assume that each of the k studies is independent of all other $k - 1$ studies, at least in the sense of employing different sampling units. There are other senses of independence, however; for example, one can think of two or more studies conducted in a given laboratory as less independent than two or more studies conducted in different laboratories. Such nonindependence can be assessed by intraclass correlations. Whether nonindependence of this type serves to increase Type I or Type II errors appears to depend in part on the relative magnitude of the Zs obtained from the studies that are correlated or too similar. If the correlated Zs are, on the average, as high (or higher) as the grand mean Z corrected for nonindependence, the combined Z one computes by treating all studies as independent will be too large. If the correlated Zs are, on the average, clearly low relative to the grand mean Z corrected for nonindependence, the combined Z one computes by treating all studies as independent will tend to be too small.

Illustration

In 1969, 94 experiments examining the effects of interpersonal self-fulfilling prophecies were summarized (Rosenthal, 1969). The mean Z of these studies was 1.014, k was 94, and Z_c for the studies combined was $9.83 = 94(1.014)/(94)^{\frac{1}{2}}$.

How many new, filed, or unretrieved studies (X) would be required to bring this very large Z down to a barely significant level ($Z = 1.645$)? By Equation 3,

$$X = (94/2.706)\,[94(1.014)^2 - 2.706] = 3{,}263.$$

One finds that 3,263 studies averaging null results ($\bar{Z} = .00$) must be crammed into file drawers before one would conclude that the overall results were due to sampling bias in the studies summarized by the reviewer. In a more recent summary of the same area of research (Rosenthal, 1976), the mean Z of 311 studies was 1.180, k was 311, and X was 49,457! Thus, nearly 50,000 unreported studies averaging a null result would have to exist somewhere before the overall results could reasonably be ascribed to sampling bias.

Discussion

There is both a sobering and a cheering lesson to be learned from careful study of Equation 3. The sobering lesson is that small numbers of studies that are not very significant, even when their combined p is significant, may well be misleading in that only a few studies filed away could change the combined significant result to a nonsignificant one. Thus, 15 studies averaging a Z of .50 have a combined p of .026; but if there were only 6 studies tucked away showing a mean Z of .00, the tolerance level for null results would be exceeded, and the significant result would become nonsignificant (i.e., $p > .05$). Or if there were 2 studies averaging a Z of 2.00, the combined p would be about .002; but uncovering 4 new studies averaging a Z of .00 would bring p into the *not significant* region.

The cheering lesson is that when the number of studies available grows large or the mean directional Z grows large, the file drawer hypothesis as a plausible rival hypothesis can be safely ruled out. If 300 studies are found to average a Z of +1.00, it would take 32,960 studies to bring the new combined p to a nonsignificant level; that many file drawers full is simply too improbable.

At the present time no firm guidelines can be given as to what constitutes an unlikely number of unretrieved or unpublished studies. For some areas of research 100 or even 500 unpublished and unretrieved studies may be a plausible state of affairs, whereas for others even 10 or 20 seems unlikely. Probably any rough and ready guide should be based partly on k so that as more studies are known it becomes more plausible that other studies in that area may be in those file drawers. Perhaps one could regard as resistant to the file drawer problem any combined results for which the tolerance level (X) reaches $5k + 10$. This seems a conservative but reasonable tolerance level; the $5k$ portion suggests that it is unlikely that the file drawers have more than five times as many studies as the reviewer, and the 10 sets the minimum number of studies that could be filed away at 15 (when $k = 1$).

It appears that more and more reviewers of research literature are estimating average effect sizes and combined ps of the studies they summarize. It would be very helpful to readers if for each combined p they presented, reviewers also gave the tolerance for future null results associated with their overall significance level.

Reference Note

1. Glass, G. V. *Primary, secondary, and meta-analysis of research.* Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1976.

References

Bakan, D. *On method.* San Francisco: Jossey-Bass, 1967.

Cochran, W. G. Some methods for strengthening the common χ^2 tests. *Biometrics,* 1954, *10,* 417–451.

Cohen, J. *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press, 1977.

Hall, J. A. Gender effects in decoding nonverbal cues. *Psychological Bulletin,* 1978, *85,* 845–857.

McNemar, Q. At random: Sense and nonsense. *American Psychologist,* 1960, *15,* 295–300.

Mosteller, F. M., & Bush, R. R. Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology: Vol. 1. Theory and method.* Cambridge, Mass.: Addison-Wesley, 1954.

Rosenthal, R. Interpersonal expectations. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic Press, 1969.

Rosenthal, R. *Experimenter effects in behavioral research* (Enlarged ed.). New York: Irvington, 1976.

Rosenthal, R. Combining results of independent studies. *Psychological Bulletin,* 1978, *85,* 185–193.

Rosenthal, R., & Gaito, J. The interpretation of levels of significance by psychological researchers. *Journal of Psychology,* 1963, *55,* 33–38.

Rosenthal, R., & Gaito, J. Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports,* 1964, *15,* 570.

Rosenthal, R., & Rosnow, R. L. *The volunteer subject.* New York: Wiley-Interscience, 1975.

Smart, R. G. The importance of negative results in psychological research. *Canadian Psychologist,* 1964, *5,* 225–232.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist,* 1977, *32,* 752–760.

Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association,* 1959, *54,* 30–34.

Received February 16, 1978 ■