under selection, I do not believe that estimates from any one of these models should be taken too seriously. Estimates from a variety of different selection models are very valuable, however, as a way to assess the sensitivity to selection effects of conclusions derived from a body of research. It may be important to know, for example, that a realistic selection model would lead to a combined estimate of treatment effect that is only half as large as that observed in published studies. This can easily happen if most of the observed effects have $p$-values only slightly smaller than the critical $p$. It is also important to know that no reasonable selection model has much effect on the combined estimate of treatment effect. This can happen when most of the observed effects have very small $p$-values. By viewing selection models as techniques for sensitivity analysis, we may exploit them more effectively in the attempt to draw scientific conclusions from collections of related research studies.

## ADDITIONAL REFERENCES

CHAMPNEY, T. F. (1983). Adjustments for selection: Publication bias in quantitative research synthesis. PhD dissertation, Univ. Chicago.

HEDGES, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *J. Ed. Statist.* **9** 61–85.

LAIRD, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73** 805–811.

LAIRD, N. M. (1982). Empirical Bayes estimates using the nonparametric maximum likelihood estimate for the prior. *J. Statist. Comput. Simulation* **15** 211–220.

LAIRD, N. M and LOUIS, T. A. (1987). Empirical Bayes estimates based on bootstrap samples. *J. Amer. Statist. Assoc.* **82** 739–750.

# Comment: Assumptions and Procedures in the File Drawer Problem

**Robert Rosenthal and Donald B. Rubin**

Interesting and important questions have been raised about the file drawer problem in the thoughtful and constructive contribution by Iyengar and Greenhouse. Our purpose here is to (a) examine the assumptions underlying the file drawer computations, (b) report some empirical estimates of retrieval bias relevant to these computations, (c) report the results of a study of retrieval bias in an early and fully documented meta-analysis and (d) comment on the framework described by Iyengar and Greenhouse and other frameworks relevant to meta-analysis.

## 1. ASSUMPTIONS UNDERLYING THE ORIGINAL FILE DRAWER COMPUTATIONS

Iyengar and Greenhouse stated that the file drawer computations (Rosenthal, 1979) are "... based upon the assumption that the unpublished studies are in fact a random sample of all studies that were done." This is, however, *not* the assumption underlying the file drawer computations proposed in Rosenthal (1979). Rather, Rosenthal (1979) explicitly assumed that (a) the null hypothesis is true (expected mean $z = 0.00$) and (b) the selection process is such that all results significant at say, .05, *two-tailed*, are published (or retrieved) whereas those that are not significant are not published (or not retrieved).

In their own assumptions underlying the file drawer computations, Iyengar and Greenhouse assume the same null hypothesis but, when critically evaluating the file drawer computations, their selection process assumption is that all results significant at say, .05, *one-tailed*, are published (or retrieved), while those that are not significant in that direction are not published (or not retrieved). In their formal models, however, Iyengar and Greenhouse assume a two-tailed selection process. Therefore, the original file drawer calculations of Rosenthal (1979) are fully consistent with all the formal models in Iyengar and Greenhouse's Section 4, which are used to illustrate their preferred maximum likelihood approach.

The Iyengar and Greenhouse file drawer calculation (based on the assumptions that the null is true but that only results significant in one direction are published) is a worst case calculation. However, it seems to be less realistic than the assumption of a two-tailed selection process because (a) early in the history of a research domain results in either direction are important news and (b) later in the history of the domain,

*Robert Rosenthal is Professor, Department of Psychology and Donald B. Rubin is Professor, Department of Statistics at Harvard University, Cambridge, Massachusetts 02138.*

when the preponderance of the evidence has supported one direction, significant reversals are often more important news than are further replications. Also, there are ample data to argue that reversals do get published. For example, of 345 retrieved studies in Rosenthal and Rubin (1978), 36% were significant (at the .05 level, one-tailed) in the predicted direction, but 3% were significant in the opposite direction.

### Minor Disagreements over Drawbacks

Iyengar and Greenhouse imply that the original file drawer computations cannot address heterogeneities in the studies retrieved. That individual users of the file drawer computations may have overlooked important heterogeneities cannot be denied. However, there is nothing to prevent stratifying studies and making file drawer computations within strata, thereby addressing simple heterogeneities.

Also, Iyengar and Greenhouse criticize the original file drawer calculations for being extremely sensitive to the selection rule (i.e., weight function). This conclusion is based on calculations like those displayed in their Table 2, however, which compare two-tailed (equation 2) and one-tailed (equation 4) selection rules. The insensitivity of Iyengar and Greenhouse's MLE procedure in Section 4 arises because only two-tailed rules are being compared. If file drawer *n*'s were compared only for these two-tailed rules, they too would be relatively insensitive to the choice of rule.

## 2. EMPIRICAL ESTIMATES OF RETRIEVAL BIAS

The extreme view of publication/retrieval bias suggests that (a) only significant studies are published and (b) only published studies can be retrieved. Consequently, in this view there are no published nonsignificant results, and no unpublished studies, significant or not, can be retrieved (Rosenthal, 1966; Sterling, 1959). Fortunately, the situation is not nearly so bleak.

For example, in the Rosenthal and Rubin (1978) meta-analysis, 61% of the 345 studies retrieved were nonsignificant. Furthermore in her meta-analysis of 11 meta-analyses, Smith (1980) reported that of 3708 results, 1285 or 35% were unpublished. Clearly, then, neither nonsignificant nor unpublished means unretrievable.

A recent study by Shadish, Montgomery and Doherty (1987) took a simple random sample of 519 possible investigators from a population of 14,002 marital/family therapy professionals. Responses concerning their research were obtained from 375 (72%) of the sample, and this yielded only 4 studies ($d = .23$) that could have been included in a meta-analysis. Shadish, Montgomery and Doherty concluded tentatively that the file drawers might contain about 149 (14,002 × 4/375) studies, which is about as many as they had retrieved for their ongoing meta-analysis (150–200).

In a survey of *all* members of a population of researchers, Sommer (1987) wrote to all 140 members of the Society for Menstrual Cycle Research. Based on a response rate of 65%, Sommer found little publication bias. Of the 73 published studies, 30% were significant in the predicted direction; of the 42 studies in the publication pipeline, 38% were significant in the predicted direction; and of the 28 studies securely filed away, 29% were significant. When only those studies were considered for which significance testing data were available, the corresponding percentages were 61%, 76% and 40%. An interesting sidelight of Sommer's study was that far and away the best predictor of publication status of the article was the productivity of the author.

Working with effect sizes rather than significance levels, Smith (1980) compared results obtained from journals, books, theses and unpublished manuscripts. Her summary of 11 meta-analyses yielded a mean effect size of $\bar{d} = .61$ for 2423 published effects compared to $\bar{d} = .51$ for 1285 unpublished effects, a difference in the direction we would predict but of a smaller magnitude than we might have expected.

In addition to work of the sort reported by Rosenthal and Rubin (1978), Shadish, Montgomery and Doherty (1987), Smith (1980) and Sommer (1987), additional studies are needed of the type described by Cochran (1963, pages 355–359) for response bias, whereby estimates of retrieval bias can be obtained by modeling the results of successive waves of survey responses. (For a summary of the procedure and for additional references see Rosenthal and Rosnow, 1975, pages 3–5.)

## 3. RETRIEVAL BIAS IN A 1969 META-ANALYSIS

### Complete versus Incomplete Retrieval

Some light may be shed on the magnitude of retrieval bias by the comparison of meta-analytic results for a research domain for which we have a complete retrieval meta-analysis and an incomplete retrieval meta-analysis.

For an earlier meta-analysis of 103 studies of interpersonal expectancy effects, studies could be divided into one group in which *all* could be retrieved because they were all conducted in a single laboratory (Rosenthal, 1969) and a second group of retrieved studies conducted elsewhere. Table 1 shows the mean $Z$ obtained for each of these two sets of studies subdivided by whether the study (a) had been published at the time of the original (1969) meta-analysis, (b) had been unpublished at the time of the meta-analysis but was published by the time of the present analysis (1987) or (c) had been unpublished at the time of the meta-analysis and remained unpublished in 1987. (The Appendix gives the individual $Z$'s in stem and leaf displays.)

Analysis of variance of the 103 studies' $Z$'s cast into the 2 × 3 table showed the interaction ($F(2, 97) = 0.49$) to be sufficiently small that the following contrasts tell the story. The comparison of retrievability yielded $t(97) = 2.92$, the comparison of studies ever published with those never published yielded $t(97) = 0.24$, the comparison of studies published at the time of the meta-analysis and those published later yielded $t(97) = 2.52$, and the comparison of studies published versus not published at the time of the meta-analysis yielded $t(97) = 2.32$.

Thus, the completely retrieved data meta-analysis yielded less significant results on average than did the incompletely retrieved data meta-analysis. That result fits our suspicions that completely retrieved data show less significant results than do less completely retrieved data. However, in this table, retrievability is fully confounded with production by a particular laboratory, which might differ in various ways from the remaining laboratories producing results bearing on the same research question.

The publication status effects are more surprising. As expected, published results were more significant than initially unpublished results. However, of the initially unpublished studies, those published eventually were *less* significant than were those never published. The impact of publication status, therefore, depends on whether we group the later published with the unpublished (as would be done at the time of the original meta-analysis) or with the published studies (as would be done if we gave unpublished studies more years to become published).

### Immediate versus Delayed Meta-Analysis

Table 2 shows that immediate meta-analysis led to a large publication status bias with $t(97) = 2.29$. However, a meta-analysis delayed to allow for eventual publication yielded essentially no publication status bias with $t(97) = 0.24$.

Information relevant to the design of a meta-analytic study is provided by the fact that the publication delays for the originally unpublished studies ranged up to 13 years with a mode of 1 year and a median of about 3 years; after 5 years, 33 of the 35 originally unpublished studies (94%) had been published.

### Proportion of Studies Found Significant

In larger meta-analyses, when many studies report only whether or not results reached some particular level of significance, it is unfortunately often necessary to replace mean $Z$'s by the proportion of inde-

TABLE 1
*Mean Z's (and their S) in two conditions of retrievability and three conditions of publication status*

| Retrievability | Publication status | | | Means |
|---|---|---|---|---|
| | Published | Published later | Never published | |
| Completely retrieved (author's lab) | $1.08^{20}$[a]<br>$(1.27)$[b] | $-0.16^{11}$<br>$(1.97)$ | $0.60^{18}$<br>$(1.15)$ | $0.63^{49}$ |
| Incompletely retrieved (other labs) | $2.60^{10}$<br>$(1.36)$ | $1.05^{24}$<br>$(1.29)$ | $1.39^{20}$<br>$(1.73)$ | $1.46^{54}$ |
| Means | $1.58^{30}$ | $0.67^{35}$ | $1.02^{38}$ | $1.06^{103}$ |

[a] Number of independent studies on which $\bar{Z}$ is based.
[b] Standard deviation of $Z$'s ($S$).

TABLE 2
*Mean Z by retrievability and publication status as classified initially or after time lapse to allow for delayed publication*

| Retrievability | Immediate meta-analysis | | Delayed meta-analysis | |
|---|---|---|---|---|
| | Published | Unpublished | Published | Unpublished |
| Completely retrieved | $1.08^{20}$[a] | $0.31^{29}$ | $0.64^{31}$ | $0.60^{18}$ |
| Incompletely retrieved | $2.60^{10}$ | $1.20^{44}$ | $1.50^{34}$ | $1.39^{20}$ |
| Means | $1.58^{30}$ | $0.86^{73}$ | $1.09^{65}$ | $1.02^{38}$ |

[a] Number of independent studies on which $\bar{Z}$ is based.

pendent studies yielding $p \leq .05$, $Z = 1.645$ (Rosenthal, 1969; Rosenthal and Rubin, 1978). Tables 3 and 4 present such results in a format analogous to that of Tables 1 and 2. The six entries of Table 1 correlated .97 with the six entries of Table 3. Similarly, Tables 2 and 4 both showed that substantial differences in results between published and unpublished studies depended upon conducting an immediate meta-analysis. In both cases a delayed meta-analysis yielded very little difference between the results of published and unpublished studies.

The effects of delaying the meta-analysis on publication bias is in great need of further research. For example, Sommer (1987) found that her 42 studies in the pipeline showed not a lower but a higher rate of significance (38%) than did either the published (30%) or unpublished (29%) results. When only those studies were considered for which sufficient information was available for significance testing, the differences were more dramatic: 76% of the pipeline studies were significant compared to 61% of the published and 40% of the unpublished studies.

### On the Null Hypothesis

There is considerable evidence to suggest that the mean $Z$ of unpublished studies is not zero but pulled into the direction of the mean $Z$ of the published studies. Thus, for just the sample in Table 1 that allowed for complete retrieval, the mean $Z$ of the initially unpublished studies was 0.31 ($n = 29$ studies) and the mean $Z$ of the never published studies was 0.60 ($n = 18$ studies).

For this meta-analysis, even these estimates are conservative. The reason is that in that early meta-analysis all $Z$'s greater than $-1.28$ and less than $+1.28$ were recorded as 0.00. As an estimate of what these 0.00 recorded $Z$'s were actually likely to be, we computed the exact $Z$ for our most recent $Z$'s falling between $Z$'s of $-1.28$ and $+1.28$. Data were available for 43 recent studies in the same research domain: the mean $Z$ was +.13, and the median $Z$ was +.22.

Of course the estimated mean $Z$ in the population of studies is not of much direct scientific interest because it depends on design considerations of the studies (e.g., sample sizes, blocking variables), which are not relevant to the science of the underlying effect. It is thus natural to try to learn something from the sampled studies about the underlying science, and Iyengar and Greenhouse's use of maximum likelihood estimation with weighted distributions addresses this objective by trying to estimate the mean effect size in the population of studies.

## 4. COMMENTS ON IYENGAR AND GREENHOUSE'S MAXIMUM LIKELIHOOD METHOD

We have already noted that the Iyengar and Greenhouse models of their Section 4 assume a two-tailed selection process, as did the original file drawer

TABLE 3

*Proportion of results significant at $Z = 1.645$ in two conditions of retrievability and three conditions of publication status*

| Retrievability | Publication status | | | Means |
|---|---|---|---|---|
| | Published | Published later | Never published | |
| Completely retrieved | .45$^{20}$[a] | .18$^{11}$ | .33$^{18}$ | .35$^{49}$ |
| Incompletely retrieved | .80$^{10}$ | .38$^{24}$ | .40$^{20}$ | .46$^{54}$ |
| Means | .57$^{30}$ | .31$^{35}$ | .37$^{38}$ | .41$^{103}$ |

[a] Number of independent studies.

TABLE 4

*Proportion of results significant at $Z = 1.645$ by retrievability and publication status as classified initially or after time lapse to allow for delayed publication*

| Retrievability | Immediate meta-analysis | | Delayed meta-analysis | |
|---|---|---|---|---|
| | Published | Unpublished | Published | Unpublished |
| Completely retrieved | .45$^{20}$[a] | .28$^{29}$ | .35$^{31}$ | .33$^{18}$ |
| Incompletely retrieved | .80$^{10}$ | .39$^{44}$ | .50$^{34}$ | .40$^{20}$ |
| Means | .57$^{30}$ | .34$^{73}$ | .43$^{65}$ | .37$^{38}$ |

[a] Number of independent studies.

computation. As they note, however, maximum likelihood estimation for those models requires far more computation than the original file drawer analysis. Of course, the output is potentially more informative because an estimate of the population mean effect size is obtained.

Accepting the general objective to estimate the mean of all studies, two comments that relate the Iyengar and Greenhouse work to other statistical ideas seem appropriate: meta-analysis as a missing data problem and finite vs. infinite population inference. A final comment addresses the scientific value of this estimand.

### The File Drawer Problem as a Missing Data Problem

There is a fairly rich statistical tradition of estimation in problems with incomplete information that can be applied to the file drawer problem by viewing study retrieval as a missing data problem. Here, in addition to the usual data from each study, $x$, there is a $1 - 0$ variable, say $R$, indicating whether that study was retrieved or not: $R = 1$ means retrieved, $R = 0$ means not retrieved. The probability specification for $R$ given $x$ is called the missing data mechanism (or the nonresponse mechanism or the retrieval mechanism) where if $R$ depends on missing $x$'s the missing data mechanism is labeled nonignorable (e.g., Rubin, 1976; Dawid and Dickey, 1977; Dempster, Laird and Rubin, 1977; Wainer, 1986; Little and Rubin, 1987). Formally, the missing data and weighted distribution approaches can be made identical here, as Iyengar and Greenhouse implicitly note when stating that they interpret $w(x)$ as the missing data mechanism.

### Finite vs. Infinite Populations of Studies

There is a potentially important difference between estimating (a) the treatment effect for the finite population of all studies that were done (retrieved and unretrieved) and (b) the treatment effect in a hypothetical infinite population of all possible studies that could have been done from which the finite population of all studies that have been done is a simple random sample. Iyengar and Greenhouse address estimation in the infinite population, which is computationally more direct than estimation in the finite population; the latter, however, is a statistically easier quantity to estimate because the observed portion of the finite population is known, and thus uncertainty of inference is restricted solely to the unobserved studies in the finite population.

Another perspective on meta-analysis, however, suggests that neither the mean effect size in the finite nor infinite populations are of basic, scientific interest.

### A New Perspective on Meta-Analysis

Rubin (1986) proposes that the current view of meta-analysis, which focuses on estimating average effects of studies, is not the most scientifically enlightening perspective. The new perspective proposed there conceptualizes meta-analysis as building and extrapolating response surfaces in an attempt to estimate "true effects," where these are defined as the effects that would be obtained in perfect hypothetical studies (e.g., randomized, infinitely large, perfectly controlled). All studies in the finite population of existing studies, or in the hypothetical infinite population from which these studies came, are flawed, and consequently in order to understand the underlying science, we should not want to summarize their typical effect size but use them to learn about that science.

Briefly the idea can be described as follows. Let $Y$ be the observed effect size for each study and classify each study by two types of characteristics: $S$, which are variables of scientific interest (e.g., strength of treatment given, sex and age of subjects, etc.) and $D$, which are design variables (sample sizes, indicators for randomized or not, laboratory indicators, etc.).

Now we can, in principle, build a response surface model for $Y$ given $(S, D)$, say $E(Y \mid S, D)$ using the observed studies, but the region where the data can fall is not of primary scientific interest because it reflects idiosyncratic choices made by investigators about values of $D$. The mean effect, $\bar{Y}$, in an idiosyncratic population of currently available, fallible studies is not of fundamental scientific interest. What is of fundamental interest is the extrapolated response surface, $Y$ given $S$ (= scientific factors) with $D$ (= design factors) fixed at $D_0$, which describes the perfect study: $E(Y \mid S, D = D_0)$. For such an objective, the overall representativeness of the studies is totally irrelevant, although representativeness given $(S, D)$ is relevant and is easier to satisfy. The task of building and extrapolating such a response surface is by no means easy, but some initial suggestions are made in Rubin (1986).

This perspective seems more damaging to the standard maximum likelihood approach to the file drawer problem, illustrated in Iyengar and Greenhouse's Section 4, than it is to the original file drawer calculation. The original file drawer calculation simply addresses the question of whether the observed batch of studies, coupled with our subjective understanding of the number of such studies that might have been done, supports the null hypothesis of no effect of the treatment—a direct and easy first step to see if there appears to be anything really going on in this domain of study. The standard maximum likelihood estimand, however, is the average effect in this population of flawed studies, which we argue is not a very scientifically relevant quantity to be estimating in any case.

## APPENDIX

*Stem and leaf display of Z's in two conditions of retrievability and three conditions of publication status*

| Retrievability | | Publication status | | |
|---|---|---|---|---|
| | | Published | Published later | Never published |
| Completely retrieved (author's lab) | 4 | | .42 | |
| | 3 | .44 | | |
| | 2 | .11, .17, .33, .33, .46 | | .81 |
| | 1 | .44, .48, .64, .69, .96 | .80 | .28, .29, .34, .48, .64, .88, .94 |
| | 0 | .00, .00, .00, .00, .00, .00, .00, .00, | .00, .00, .00, .00, .00 | .00, .00, .00, .00, .00, .00, .00, .00 |
| | −0 | | | |
| | −1 | .52 | .99, .50 | .51, .28 |
| | −2 | | .33, .17 | |
| Incompletely retrieved (other labs) | 5 | | | .24, .38 |
| | 4 | .67 | | |
| | 3 | .25, .25, .37, .96, | .70, .89 | .29 |
| | 2 | .05, .10 | .02, .14, .27 | .01, .33, .61 |
| | 1 | .48, .88 | .34, .44, .46, .55, .64, .64, .65, .80 | .60, .60, .83, .95 |
| | 0 | .00 | .00, .00, .00, .00, .00, .00, .00, .00, .00, .00 | .00, .00, .00, .00, .00, .00, .00, .00, .00, .00 |
| | −0 | | | |
| | −1 | | .45 | |

## ACKNOWLEDGMENT

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

## ADDITIONAL REFERENCES

Cochran, W. G. (1963). *Sampling Techniques*, 2nd ed. Wiley, New York.

Dawid, A. P. and Dickey, J. M. (1977). Likelihood and Bayesian inference from selectively reported data. *J. Amer. Statist. Assoc.* **72** 845–850.

Dempster, A. P., Laird, N. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research.* Appleton-Century-Crofts, New York.

Rosenthal, R. (1969). Interpersonal expectations. In *Artifact in Behavioral Research* (R. Rosenthal and R. L. Rosnow, eds.). Academic, New York.

Rosenthal, R. and Rosnow, R. L. (1975). *The Volunteer Subject.* Wiley, New York.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

Rubin, D. B. (1986). A new perspective on meta-analysis. Paper presented at the National Research Council Workshop on the Future of Meta-Analysis, Hedgesville, W. Va.

Shadish, W. R., Jr., Montgomery, L. M. and Doherty, M. (1987). How many studies are in the file drawer? An empirical estimate. Paper presented at the Meeting of the American Evaluation Association, Boston.

Smith, M. L. (1980). Publication bias and meta-analysis. *Evaluation Ed.* **4** 22–24.

Sommer, B. (1987). The file drawer effect and publication rates in menstrual cycle research. *Psychol. Women Quart.* **11** 233–242.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Amer. Statist. Assoc.* **54** 30–34.

Wainer, H. (ed.) (1986). *Drawing Inferences from Self-Selected Samples.* Springer, New York.