# Comment

**William DuMouchel**

## 1. THE SELECTION MODEL APPROACH TO THE FILE DRAWER PROBLEM

Professors Iyengar and Greenhouse provide a nice introduction to the file drawer problem and then criticize and improve the so-called "fail-safe" method of Rosenthal (1979), as well as its modification in Orwin (1983). In doing so, they provide a strong argument that Orwin's modification, based on averaging effect sizes, is to be preferred to the Rosenthal version, based on computing a joint $p$-value for a set of studies.

But the major purpose of their paper is to present and advocate another approach, an extension of that of Hedges and Olkin (1985), in which the probability of selection bias for each study is explicitly incorporated into the likelihood function. Iyengar and Greenhouse provide an example meta-analysis of the ten studies summarized in Table 4, to illustrate the use of their method. Using either of two proposed parametric models of selection bias, they jointly estimate the degree of selection bias and the common effect size in the ten studies. The resulting inferences are claimed to be more complete, useful and robust than those from the fail-safe method.

Although I am inclined to grant them those modest claims, I would have liked to have seen a deeper discussion of their own example. On page 13 the authors state "Perhaps the most interesting feature of the log likelihood contours is their width . . . this meta-analysis is not very informative for $\theta$." I disagree with both clauses. To take the latter clause first, a standard error of .05 for $\theta$ seems quite an achievement for estimation of any effect size, and I suspect that most social scientists reviewing the data in Table 4, where estimates range from $-.58$ to 1.05, would call it unduly optimistic. Second, that feature of the figures which startled me the most, and which is also reflected in the covariance matrices of Table 5, is that there is hardly any correlation between the estimates of $\theta$ and $\beta$ (or of $\theta$ and $\gamma$). If we work with the normal approximation to the joint posterior distribution of $(\theta, \beta)$ as shown in Table 5, the correlation between the two parameters is about $-0.1$. Thus, knowledge that $\beta = .1$ would lead one to predict $\theta = 0.036$, although increasing $\beta$ by four standard deviations to 2.5 leads to the prediction of $\theta = 0.016$. Large changes in the assumed size of the selection bias lead to trivial changes in the estimate of the parameter of interest. Does this mean that the file drawer problem is not a problem at all?

Looking at their Figure 1 more closely, it can be seen that there is a lot more information in the likelihood function for $\theta$ (the contours are more closely spaced in the vertical direction) when $\beta = 2.5$ than when $\beta = 0.1$; does this mean that selection bias helps us estimate $\theta$? One can get some insight into this puzzling behavior by studying equation (6). Because the numerator of the expression for $L(\theta, w)$ factors into a function of $\theta$ alone and a function of $w$ (i.e., $\beta$) alone, all the dependence between the two parameters in the joint likelihood must come from the term in the denominator of (6). Now this term does not even depend on the data $(t_1, \ldots, t_{10})$, but only on the sample sizes $N_i$ and on the assumed forms of $f(t; \theta)$ and of $w(t)$. Thus, inferences drawn about the relationship between $\theta$ and $\beta$, based on that likeliood function, may depend more on prior assumptions than on the data.

## 2. THE ASSUMPTION OF HOMOGENEOUS EFFECT SIZES

The authors state that "for illustrative purposes (and following Hedges and Olkin) we assume that all studies are estimating the same effect size." I hereby propose that statisticians never recommend for general use any method of meta-analysis which does not include, somewhere in the model, a parameter or component of variance for between study variability. This admittedly extreme proposal would rule out recommendation of the inverse normal method of combining one-tailed significance tests, and thus the Rosenthal fail-safe sample size procedure, because, contrary to the implication of the authors in Section 3, one needs to assume that *every* effect size is zero, not just that the mean effect size is zero, in order to have each $p_i$ uniformly distributed on [0, 1]. The Orwin procedure based on average effect sizes does not seem to need this assumption, providing another reason to prefer it to Rosenthal's version.

Figure 1 graphs the ten estimates from Table 4 with approximate 95% error bars based on plus or minus two standard errors. A glance at Figure 1 makes it

*William DuMouchel is Chief Statistician at BBN Software Products Corporation, 10 Fawcett Street, Cambridge, Massachusetts 02238.*

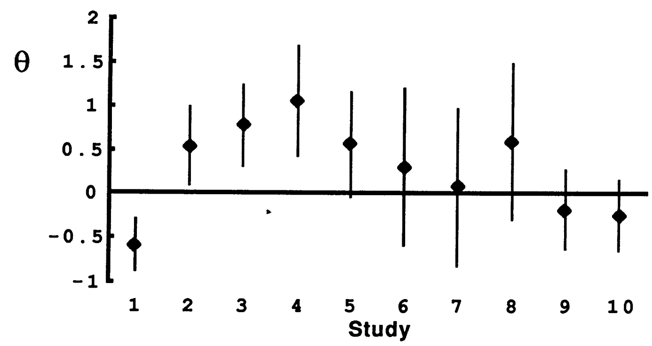


FIG. 1. *Estimates (±2 SE) for each study from Table 4.*

clear that the hypothesis of a common value of $\theta$ for every study is ruled out. Studies 2, 3 and 4 definitely have higher effect sizes than average, studies 1, 9 and 10 are definitely lower than average, although the other four studies fall somewhere in the middle. That granted, just what does the subsequent analysis, based on the assumption of equal $\theta$'s, illustrate? It is hard to tell whether the features of the method being discussed, such as the behavior of the likelihood contours, are typical or just due to a poorly fitting model. Furthermore, the standard errors presented in Section 4 depend on the model being appropriate.

Can the authors' selection model be extended to handle with case where the ten "true values" of $\theta$ have been drawn from a superpopulation with mean $\mu$ and variance $\sigma^2$? Perhaps it can, especially if the statistician has, and is willing to use, prior information about the distribution of true values and about the mechanisms governing the selection bias. In their conclusion, the authors raise the question of design issues for meta-analyses. What possible design issues can arise if the meta-analysis uses such a simplified model? Once one admits a component of variance for between study variation, the trade-off between making many smaller studies or fewer large studies begins to get interesting. If, in addition, one uses other characteristics differentiating the studies to build a hierarchical prior distribution for the $\theta$'s, then design considerations can become paramount, as discussed in DuMouchel and Harris (1983) and DuMouchel and Groër (1987).

## ADDITIONAL REFERENCE

DUMOUCHEL, W. and GROËR, P. (1987). A Bayesian methodology for scaling radiation studies from animals to man. Presented at the 26th Hanford Life Sciences Symposium, October 1987. *Health Physics*. To appear.

# Rejoinder

**Satish Iyengar and Joel B. Greenhouse**

Our objectives in writing this paper were to illustrate a practical application of selection models as a technique for sensitivity analysis in meta-analysis, to develop further statistical methodology for the file drawer problem and to identify statistical and practical issues related to the theory and practice of meta-analysis. We are indebted to the discussants on several accounts. Each of them has made fundamental contributions to the selection model or meta-analysis literature, and the roots of this paper are found in these earlier works. Furthermore, in their comments they have suggested modifications and alternative approaches that are likely to improve the methods discussed as well as the general practice of meta-analysis.

Professor Hedges and Professors Rosenthal and Rubin suggest that the issue of publication bias is overemphasized. Hedges believes that the related problem of "reporting bias" where studies test many hypotheses and report sufficient statistics only for results that achieve statistical significance is more widespread. Rosenthal and Rubin point out that the usual file drawer problem portrays a rather extreme view of publication bias and in fact empirical research shows that "neither nonsignificant nor unpublished means unretrievable." Although both of these points are well taken, nevertheless, Dickersin, Chan, Chalmers, Sacks and Smith (1987) report that "the results of published RCTs (randomized clinical trials) are more likely to favor the new therapy than are the results of unpublished RCTs . . . " and conclude that " . . . it seems likely that bias against the publication of 'negative' results does exist." As we note in our paper, with the general interpretation of the weight function as a model for the selection mechanism, both reporting and retrieval bias can be treated as special cases of the general methodology. Finally, we hope that as authors and editors of journals begin to adopt guidelines for reporting statistical studies, such as those suggested by Bailar (1986), the problems of reporting bias might diminish.

The empirical results concerning publication bias presented by Rosenthal and Rubin are interesting but