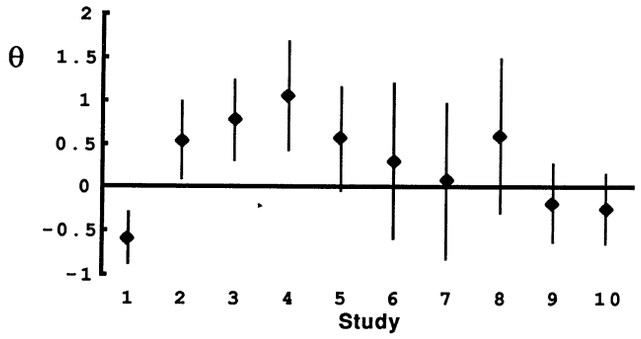


FIG. 1. *Estimates* (±2 *SE*) *for each study from Table* 4.

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

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science*. www.jstor.org

raise many questions that are in themselves worthy of a full discussion. We are concerned that the analysis of the data in Rosenthal and Rubin's Table 1 of completely retrieved versus incompletely retrieved studies is inextricably confounded with the fact that the completely retrieved studies all come from Professor Rosenthal's laboratory. Hence, the observed direction and differences in effect sizes can be explained by the carefulness of the research for which Professor Rosenthal is noted and not necessarily by retrieval status. Their second demonstration that "a meta-analysis delayed to allow for eventual publication" will eliminate publication status bias is puzzling. Surely Rosenthal and Rubin do not mean to suggest that by delaying a meta-analysis the problems of publication/retrieval bias would be resolved. Any conclusion based on such an analysis must take into account the studies completed after the initial meta-analysis and their subsequent publication status. Furthermore, even if we allow for a weight function that does not change over time, the analysis of the data in Rosenthal and Rubin's Table 2 is analogous to the two-stage sampling procedure described by Glynn, Laird and Rubin (1986) for nonresponders in a sample survey and should be analyzed accordingly (see (2) in the comments by Laird, Patil and Taillie).

All of the discussants are unhappy with our use of a model that does not account for between study variability. So are we. In our attempt to simplify the introduction of weighted distributions to meta-analysis we missed an opportunity to present the fuller analysis. We have discussed the use of hierarchical models for combining study results elsewhere (Greenhouse, Fromm, Iyengar, Dew, Holland and Kass, 1986) (although without accounting for selection bias), emphasizing the need, as our discussants do, to take into account the heterogeneity among studies. We therefore heartily endorse Professor DuMouchel's proposal.

Professor Bayarri's thoughtful formulation of the fully Bayesian analysis unifies the estimation of the fail-safe sample size and the effect size in the presence of selection and directly incorporates Professor Rao's suggestion of dealing with heterogeneity by putting a prior distribution on $\theta$. In particular, an appealing feature of this approach is the opportunity to study how inferences about $\theta$ depend on $N$. By writing the marginal posterior distribution of $\theta$ as

$$(1) \qquad p(\theta \mid t, k) = \sum_N p(\theta \mid N, t, k) p(N \mid t, k)$$

we can investigate how $p(\theta \mid N, t, k)$ changes with $N$. Thus, in cases where $p(\theta \mid N, t, k)$ is very sensitive to changes in $N$, it would be important to examine carefully the distribution $p(N \mid t, k)$, which summarizes the information about $N$ in the light of the data and prior knowledge.

With respect to our example, Professors Laird, Patil and Taillie wonder whether the lack of sensitivity to the choice of weight function is due to the symmetry of the weight function about $t = 0$ and to the fact that both weight functions give uniform weight to the tails. Hedges also wonders whether our choice of weight functions is realistic. To address these concerns and to further investigate the sensitivity of our inferences about $\theta$ to our assumptions about the weight function, we consider a third weight function, which is asymmetric:

$$(2) \quad w_3(x; \alpha, \beta) = \begin{cases} 1 & \text{for } x > t(q, .05), \\ e^{-\alpha} & \text{for } |x| \le t(q, .05), \\ e^{-\beta} & \text{for } x \le -t(q, .05), \end{cases}$$

where $t(q, .05)$ is defined as in $w_1$ and $w_2$ of Section 4. The maximum likelihood estimates of the parameters are $(\hat{\alpha}, \hat{\beta}, \hat{\theta}) = (3.6, 2.5, -0.075)$, with standard errors (1.08, 2.03, 0.092), respectively, and correlation matrix

$$(3) \quad \operatorname{corr}(\hat{\alpha}, \hat{\beta}, \hat{\theta}) = \begin{pmatrix} 1.00 & 0.77 & -0.76 \\ 0.77 & 1.00 & -0.81 \\ -0.76 & -0.81 & 1.00 \end{pmatrix}.$$

Although this may be a special case, we see that with the asymmetric weight function (i) inferences about $\theta$ are largely unaffected and (ii) the estimate of $\theta$ is highly correlated with the estimates of the selection parameters. The choice of a weight function is, of course, a difficult one and we agree with Hedges that the ability to analyze several weight functions as we have done here is a powerful technique.

Professor DuMouchel and Professors Laird, Patil and Taillie note that a remarkable feature of our example is that the estimates of effect size ($\hat{\theta}$) and the weight function parameters ($\hat{\beta}$ or $\hat{\gamma}$) are nearly uncorrelated. Laird, Patil and Taillie conjecture that this is due to a combination of the use of a symmetric weight function and the absence of an overall effect. Partial support for this claim comes from an investigation of the following similar problem: let the underlying density be normal with mean $\theta$ and variance 1, let the weight function be

$$(4) \qquad w(x; \beta) = \begin{cases} e^{-\beta} & \text{for } |x| \le 1.96, \\ 1 & \text{for } |x| > 1.96 \end{cases}$$

and consider the expected information matrix $I(\theta, \beta)$. Routine calculations show that for $\theta = 0$, $I(\theta, \beta)$ is diagonal and the correlation is zero. As $\theta$ increases from zero, the correlation does initially increase, but then tends to zero as $\theta$ tends to infinity. The same analysis with an asymmetric weight function shows that even when $\theta = 0$ the correlation can be substantial.

In all due respect to Professors Rosenthal and Rubin, Rosenthal (1979) never "explicitly assumed" a

two-tailed selection process. We agree that the one-sided selection is a "worst case calculation," and that the two-sided selection with equal weights on both sides is consistent with our models in Section 4. However, even this latter file drawer calculation is quite sensitive to departures from these assumptions. In fact, our argument of Section 3 shows that the fail-safe sample size will in general be much smaller for unequal weights (so that the weighted null mean is non-zero). Notice that on the other hand, the maximum likelihood procedure with the asymmetric weight function above gave results similar to those given by $w_1$ and $w_2$.

Rosenthal and Rubin propose a new perspective on meta-analysis based on "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." It follows from this perspective that not only are all existing statistical methods for meta-analysis flawed but also the very studies upon which every meta-analysis is based. It seems to us that in so far as a meta-analysis incorporates methods for distinguishing good quality studies from the bad, the objectives of the Rosenthal and Rubin new perspective will be achieved. In the meantime, as they indicate, the task of building and extrapolating a response surface will not be easy in practice.

Cochran (1954) makes the point that "the combination of individual estimates is not a routine matter, but requires clear thinking about both the nature of the data and the function of a combined estimate." We have been most fortunate to have benefited from the clear thinking and insightful comments of the discussants for which we thank them.

## ADDITIONAL REFERENCES

BAILAR, J. C., III (1986). Reporting statistical studies in clinical journals. In *Medical Uses of Statistics* (J. C. Bailar, III and F. Mosteller, eds.) 261–271. NEJM Books, Waltham, Mass.

COCHRAN, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10** 101–129.

DICKERSIN, K., CHAN, S., CHALMERS, T., SACKS, H. and SMITH, H. (1987). Publication bias and clinical trials. *Controlled Clin. Trials* **8** 343–353.

GLYNN, R. J., LAIRD, N. M. and RUBIN, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples* (H. Wainer, ed.). Springer, New York.