Comment

Larry V. Hedges

The paper by Iyengar and Greenhouse represents an important advance in more realistic modeling of selection bias in meta-analysis. Previous work on estimation of effect size under selection bias (e.g., Hedges, 1984; Champney, 1983) has been restricted to the case of simple truncation. By moving beyond the simple truncation model to one in which the probability of observing an effect increases monotonically with effect size, they have posed a much more realistic model of selection bias. My comments suggest other ways that these models might be made more realistic. I organize them around four areas: origins of selection bias, the choice of weighting functions, models for combining estimates and uses of estimation procedures incorporating selection bias.

ORIGINS OF SELECTION BIAS

One form of selection bias is publication bias that connotes the effect of journal editorial policies that result in rejection of manuscripts that do not obtain results that are statistically significant (at conventional levels). One might also imagine that sophisticated authors who did not obtain statistical significance would spare themselves the embarrassment of a rejection and would not submit their manuscripts for publication. Although unpublished studies languishing in researchers' file drawers occasionally have been found, I doubt that this is the major source of missing data on effect magnitudes. I believe that the related problem that I have called reporting bias is more widespread. Although it is often convenient to idealize studies as having only one treatment contrast and one outcome, virtually all studies are more complex, involving several hypothesis tests. I have found that studies in the social sciences tend to test many hypotheses and report sufficient statistics only for results that achieve statistical significance. The statistics they report for results that are not statistically significant are usually less complete and may consist only of a notation that the results in question were "not significant." Because the estimate of effect cannot be calculated from the statistics reported when results are not statistically significant, reporting bias has effects on estimation that are similar to those of truncation induced by an editorial process that will not publish papers whose results are not statistically significant. When the number of studies that did not achieve statistical significance is known, as in the reporting bias model, it may be more appropriate to utilize this information by considering the situation as a type I censoring problem of estimating the mean when the number of censored observations is known. Estimation using the EM algorithm would be quite feasible under this censoring model.

THE RELATIONSHIP BETWEEN *p*-VALUES AND THE PROBABILITY THAT RESULTS ARE REPORTED

More realistic models of the selection process require careful consideration of the relationship between estimates (or p-values) and the probability that a result is observed. The use of a parametric family of weight functions with parameters to be estimated from the data could result in a much more realistic model of the selection process than the step functions used for weights by previous authors.

The families of weight functions considered by Iyengar and Greenhouse, however, are probably unrealistic because they imply that significant results are observed with probability one. Well designed studies with essentially zero effect are sometimes published and I can cite some examples of poorly designed studies that obtained very small p-values but were not accepted for publication. The general principle is that when p-values are either very small or very large, the decision whether to report or publish is based primarily on other factors than the p-value (e.g., design of the study). When the p-value is intermediate, the decision to publish (report) may be greatly influenced by the p-value. Thus, it may be more realistic to model the weight function relating effect size to probability of observation by an s-shaped curve like a logistic function rather than a power function or an exponential. For example: $e^{\alpha+\beta|\Theta|}/(1 + e^{\alpha+\beta|\Theta|})$, where Θ is the effect magnitude, α is a parameter that sets the probability that a result is observed even when $\Theta = 0$ and β is a constant effectively determining the slope of the curve near its inflection point. Even if the parameters of the weight function prove difficult to estimate precisely, a function with more realistic functional form seems preferable to a weight function whose form fails to capture an important aspect of the selection process.

PUBLICATION BIAS AND MODELS FOR COMBINING ESTIMATES

Two slightly different classes of statistical models have been used in combining estimates from several

Larry V. Hedges is Associate Professor and Chairman of the Measurement, Evaluation and Statistical Analysis (MESA) Program, Department of Education, University of Chicago, 5835 South Kimbark Avenue, Chicago, Illinois 60637.

studies, differing in the way that they conceptualize between study differences in effects. Let $t_1, \cdots, t_k$ be the sample effect estimates from k studies and let $\Theta_1, \cdots, \Theta_k$ be the corresponding population parameters. Fixed effects models treat $\Theta_1, \cdots, \Theta_k$ as fixed but unknown constants. Iyengar and Greenhouse consider the effects of publication bias on estimates in a fixed effect model with $\Theta_1 = \cdots = \Theta_k = \Theta$.

Although this model of homogeneity of results across studies is convenient and a logical starting point for analyses, meta-analyses often find evidence of substantial between study heterogeneity in effects. It would be highly desirable to incorporate a mechanism for evaluating the plausibility of the assumption that $\Theta_1 = \cdots = \Theta_k$. Note, for example, that the estimates presented in Table 4 exhibit considerable heterogeneity. Figure 1 depicts the point estimates and 95% confidence intervals for Θ derived from each study individually (that is, based on the conditional distribution of t_i given Θ_i). If the estimates are treated as unselected, a conventional test for heterogeneity (see Hedges and Olkin, 1985, page 123) would reject the hypothesis that $\Theta_1 = \cdots = \Theta_k$ at the .001 level of significance.

Random effects models for combining study results treat the "population effects" $\Theta_1, \cdots, \Theta_k$ from k studies as if they were sampled from a hyperpopulation. The rationale for random effects models is that different studies usually implement the treatment in a slightly different fashion. Thus, it is most sensible to conceptualize the k studies as having a sample of k treatment implementations drawn from a universe of possible treatment implementations rather than as k studies implementing the treatment identically. If variations in treatment lead to variations in treatment effect it is therefore sensible to model treatment effects as sampled from a distribution. In random effects models, the combination problem is to estimate the parameters (typically the mean $\bar{\Theta}$ and variance σ^2_Θ) of the distribution from which $\Theta_1, \cdots, \Theta_k$ were drawn. Champney (1983) considered the effects of publication bias in random effects models in which the random effects are normally distributed and the weight function was a step function. His work suggested that publication bias may have substantial effects on estimation of the variance component σ^2_Θ even when the estimate of the mean $\bar{\Theta}$ is not strongly affected. It would be interesting and relatively straightforward to study the effects of publication bias on estimates in random effects models when more sophisticated and realistic weight functions (like those of Iyengar and Greenhouse) are used.

One thrust of recent research on random effects models in meta-analysis is the use of mixture modeling in which the distribution of the random effects is estimated from the data (Laird, 1978, 1982; Laird and Louis, 1987). It seems quite likely that estimates of the distribution of the random effects (the mixing distribution) are sensitive to selection bias. Further work to elucidate the effects of selection on estimates of the distribution of treatment effects would be an important contribution.

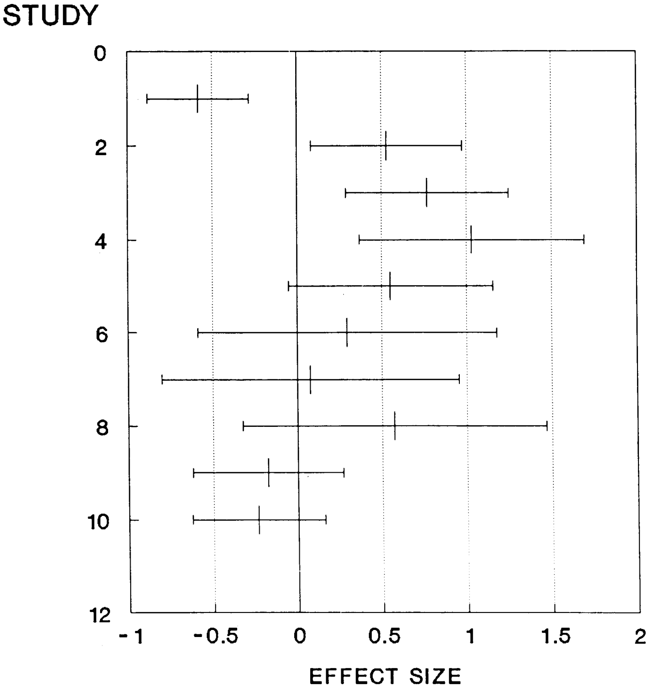

FIG. 1. *Point estimates and 95% confidence intervals for $\Theta_1, \ldots, \Theta_{10}$ based on $t_1, \ldots, t_{10}$ from Table 4.*

USES OF ESTIMATION BASED ON SELECTION MODELS

The selection models operating are likely to depend on the conventions of research and research reporting. Consequently, the choice of particular selection models for meta-analyses is a murky business. Different (but reasonable) choices for weight functions may give different estimates of the combined effect. There is, however, very little empirical evidence to guide the choice of weight functions. Such evidence might be obtained by making use of comprehensive registries of studies. By obtaining estimates and p-values from unpublished studies located through the registry, it should be possible to suggest reasonable choices for weight functions in selection models. However, because reporting and publication practices involve social (e.g., scientific) conventions, it is not obvious that those choices of weight functions would generalize from one domain of research to another.

Although I am enthusiastic about the development of more varied and realistic models for estimation 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.