Statistical Science
1992, Vol. 7, No. 2, 246–255

Modeling Publication Selection Effects in Meta-Analysis

Larry V. Hedges

Abstract. Publication selection effects arise in meta-analysis when the effect magnitude estimates are observed in (available from) only a subset of the studies that were actually conducted and the probability that an estimate is observed is related to the size of that estimate. Such selection effects can lead to substantial bias in estimates of effect magnitude. Research on the selection process suggests that much of the selection occurs because researchers, reviewers and editors view the results of studies as more conclusive when they are more highly statistically significant. This suggests a model of the selection process that depends on effect magnitude via the p-value or significance level. A model of the selection process involving a step function relating the p-value to the probability of selection is introduced in the context of a random effects model for meta-analysis. The model permits estimation of a weight function representing selection along the mean and variance of effects. Some ideas for graphical procedures and a test for publication selection are also introduced. The method is then applied to a meta-analysis of test validity studies.

Key words and phrases: Publication bias, selection models, file-drawer problem, meta-analysis, random effects models, weight function models.

Combining information across replicated research studies is a fundamental scientific activity. In the last 15 years there has been a growing appreciation in education, psychology and the medical sciences that systematic procedures for combining evidence are critical to ensure the validity of conclusions drawn from such evidence (Cooper, 1984; Glass, McGaw and Smith, 1981; Rosenthal, 1984; Peto, 1987). One aspect of these systematic procedures is the use of explicit statistical methods to combine estimates across studies—often known as meta-analysis (e.g., Hedges and Olkin, 1985).

Statistical methods for meta-analysis usually are estimation procedures. They are designed to have desirable properties (e.g., consistency and asymptotic efficiency) when the sample of estimates available is an unselected sample of those that could have been obtained from experiments like those conducted in the studies whose results are to be combined. The seemingly innocuous assumption that the sample of results (estimates of effect magnitude) available for combination can be considered a representative sample of an unselected population of such estimates is often highly debatable, and sometimes is demonstrably false. If this assumption is not met, statistical methods based on the assumption can produce misleading results.

THE PROBLEM OF PUBLICATION SELECTION BIAS IN META-ANALYSIS

Publication selection bias (and any other form of selection) is difficult to document because of the nature of the phenomenon. To do so we need information about the missing estimates which is usually unavailable. There is, however, some reasonably direct and other indirect evidence about the existence of publication bias.

Failure to Report Nonsignificant Results

One readily demonstrable source of bias in reported estimates of effect magnitude stems from the failure to report the details of statistical analyses when mean differences are not statistically significant. Calculation of an effect magnitude estimate requires the values of sample statistics. Authors sometimes fail to report either the test statistics or descriptive statistics but simply report "no significant difference." This tendency to report the details of statistical analyses only when significant results are obtained is part of what has

Larry V. Hedges is Professor of Education at The University of Chicago, 5835 South Kimbark Avenue, Chicago, Illinois 60637.

been labeled "prejudice against the null hypothesis." Although statisticians are unlikely to condone such conditional reporting of statistical results, the practice is prevalent in psychological and medical research. Apparently, statistical indoctrination about null hypothesis tests has led many researchers to believe that it is incorrect to interpret the results of research when the null hypothesis cannot be rejected. One consequence of this practice is that the data needed to calculate effect size estimates are sometimes selectively omitted in research reports when insignificant results are obtained. For example, in meta-analyses by Eagly and Carli (1981) and Strube (1981), as many as 40% of the potential effect size estimates could not be calculated because of incomplete reporting of nonsignificant results.

Although the selective omission in published papers of results that are statistically nonsignificant is relatively easy to document, this practice no doubt leads to failure to submit for publication entire research reports that examine one principal hypothesis test. For example, while 60% of psychological researchers in one survey indicated that they would submit a research report for publication if its results were statistically significant, only 6% indicated that they would do so if the results were not statistically significant (Greenwald, 1975). Similar results were obtained in a survey of medical researchers who were authors of published studies (Dickersin et al., 1987). They identified 125 studies that were completed but not published. The vast majority (82%) of these unpublished studies were not even submitted for publication. The lack of a beneficial medical effect (i.e., nonsignificance), cited in 34% of the cases, was the most frequent reason for failure to publish.

Editorial Policy

There is evidence that editorial policy in many areas *intends* to discourage the publication of research reports that yield statistically insignificant results (Bakan, 1966; Sidman, 1960). Perhaps the clearest statement of such a policy was made by the editor of the *Journal of Experimental Psychology*, one of the most prestigious journals in psychology:

> In editing the journal there has been a strong reluctance to accept and publish results related to the principle concern of the research when those results were significant at the 0.05 level, whether by one- or two-sided test. This has not implied a slavish worship of the 0.01 or any other level, as some critics may have implied. Rather, it reflects the belief that it is the responsibility of the investigator in a science to reveal his effect in such a way that no reasonable man would be in a position to discredit the results by saying that they were the product of the way the ball bounced. At least, it was believed that such findings do not deserve a place in an archival journal, even though they may be proper fare for symposia, scientific meetings, and dittoed handouts (Melton, 1962).

Melton certainly implies that articles with more highly significant results are more likely to be accepted. Although unusually explicit in his views, Melton's position on the importance of statistical significance is widely shared in education, psychology and the behavioral sciences. This is verified both by empirical studies (e.g., Coursol and Wagner, 1986) and by surveys of journal editors and reviewers. For example, a recent survey of reviewers for major journals in psychology revealed that statistical significance of results is an important criterion used in evaluating studies for publication (Greenwald, 1975).

Proportion of Significant Results in the Published Literature

The published research literature in the social and behavioral sciences contains a very high proportion of statistically significant results. A survey of three psychology journals by Sterling (1959) and a more extensive survey of psychology journals by Bozarth and Roberts (1972) showed that 97% and 94%, respectively, of the articles examined that used statistics rejected the statistical null hypothesis at the $\alpha = 0.05$ significance level. If the studies examined by Sterling, and Bozarth and Roberts are representative of all studies conducted in psychology, their results imply either a large number of Type I errors, an average power in excess of 0.90, or some combination of these possibilities.

The overwhelming proportion of articles rejecting the null hypothesis in these surveys is indeed remarkable, given the results of studies of the power of statistical tests in psychological research (Chase and Chase, 1976; Cohen, 1962). These surveys of statistical power suggest that the average power of statistical tests in psychological research is between 0.25 and 0.85, depending on the assumed magnitude of effects. Therefore, the surveys of statistical power suggest that between 25% and 85% of the studies in psychology journals would be expected to yield statistically significant results. The larger than expected proportion of significant results in the published literature may be evidence of selection bias favoring publication of statistically significant results.

Comparison of Published and Unpublished Studies

One direct means of studying publication bias is to compare samples of published and unpublished studies that presumably estimate the same treatment effect. Because statistically significant results tend to be associated with larger effects, a tendency of unpublished studies to have a smaller effect magnitude than published studies would suggest the presence of publication bias. Several authors have found just such a pattern of smaller effects in published studies (Smith, 1980; Devine and Cook, 1983; Dawes, Landman and Williams, 1984; White, 1982).

Comparisons of Studies with Different Sample Sizes

Another method of determining the existence of publication bias is to compare the distributions of effect magnitude estimates obtained in studies with different sample sizes. If sample size is unrelated to the true effect magnitude (which seems quite plausible) and there is no publication bias, then a plot of the sample size versus effect magnitude estimates should resemble a funnel – estimates with smaller sample size should be "spread out" over a wider range, but the center of the distribution should be the same for each sample size (Light and Pillemer, 1984). If there is publication bias, the funnel is distorted with fewer small effect estimates for small sample sizes and a tendency for the center of the funnel to be tilted (since publication bias distorts the mean estimate most for small sample sizes). Applications of this technique to data on the effects of psychoeducational care (Light and Pillemer, 1984) and to data from clinical trials (Berlin, Begg and Louis, 1989) also suggest the existence of publication bias.

Consequences of Publication Bias

If effect magnitude estimates corresponding to statistically insignificant results are less likely to be sampled (or to be available for sampling), then the sample of effect size estimates will be biased. Because test statistics are usually monotonically related to effect magnitude estimates, studies that produce statistically significant results tend to have effect sizes that are larger in absolute magnitude. Thus, for positive population effect sizes, overrepresentation of studies producing significant differences will tend to bias the sample toward larger effect sizes. Hence each effect size estimate or average of estimates from such a sample will overestimate the absolute magnitude of the population effect size.

Lane and Dunlap (1978) studied an extreme form of bias toward significant results: the situation in which only experiments yielding statistically significant results are observed. They simulated the results of a large number of two-group experiments and selected for further study experiments that yielded statistically significant mean differences. Hedges (1984) studied the same problem analytically. As expected, both studies found that the mean difference estimated from experiments yielding significant results overestimated the population mean difference.

Hedges (1984) and Hedges and Olkin (1985) studied the effects of the same extreme model of publication bias on estimates of the standardized mean difference and Hedges (1989) studied its effects on estimates of the mean and variance. They demonstrated that the bias induced by selection is a function of both sample size and the true effect magnitude and that the bias can be very large when either the true effect magnitude or the sample size is small. Begg and Berlin (1988) obtained similar results under a slightly different model of publication selection.

While these theoretical analyses have tended to examine rather extreme models of publication bias in which no statistically insignificant results are reported, they demonstrate that publication bias *can* have substantial effects.

CORRECTING FOR PUBLICATION BIAS

Research on corrections for publication bias has taken one of three principal approaches. The first approach is to investigate the sensitivity of the conclusions obtained by meta-analytic procedures to the possibility that there are other unpublished results lurking "in the file drawer" of researchers (Rosenthal, 1979; Orwin, 1983; Iyengar and Greenhouse, 1988). A second approach is the use of invariant sampling frames, such as registries of studies. By restricting attention to studies that are registered prior to the collection of data, biases on study results due to selection can be eliminated (Simes, 1986a, b; Begg and Berlin, 1988). The third approach, the one taken by this paper, is to model the selection process and develop estimation procedures that take that selection process into account.

Hedges (1984) presented methods for estimation of the standardized mean difference under a truncation model of selection. In this selection model, estimates are observed (selected) only if they are statistically significant at the $\alpha = 0.05$ level. Champney (1983) treated estimation in random effects analyses under the same selection model. Iyengar and Greenhouse (1988) extended earlier models for selection by introducing the general use of weighted distributions (Rao, 1965, 1985) to model selection effects. This perspective is a generalization because it includes earlier work on selection modeling as a special case in which the weight function is an indicator function (i.e., a function taking the value 1 if the result is statistically significant and 0 if it is not). Iyengar and Greenhouse examined estimation under two weight functions with prespecified parametric form.

In this paper we generalize the weight function method of Iyengar and Greenhouse (1988). Instead of specifying a parametric form for the weight function, we approximate the weight function by a step function and estimate the value of the weight function on each interval. The method described in this paper is quite similar to that described by Dear and Begg (1992). It differs only in that we rely on external information (from psychological studies of the interpretation of p-values) to specify the location of the discontinuities in the weight function while they estimate the points of discontinuity directly from the data.

MODEL AND NOTATION

A Model for Study Results in the Absence of Selection

Let $X_1, \ldots, X_n$ be variables such that

$$X_i \sim N(\delta, \sigma_i^2)$$

where σ_i^2 is known and δ is an unknown parameter distributed such that

$$\delta \sim N(\Delta, \sigma^2),$$

where Δ and σ^2 are both unknown. Thus

$$X_i \sim N(\Delta, \sigma_i^2 + \sigma^2). \tag{1}$$

The observed statistic X_i is used in study i to test the hypothesis H_0: $\delta_i = 0$ using the test statistic

$$Z_i = |X_i|/\sigma_i$$

and the (two-tailed) p-value associated with this test is

$$p_i = 1 - \Phi(Z_i) + \Phi(-Z_i) = 2\Phi(-Z_i). \tag{2}$$

A Model for Selection Bias

Selection bias arises as a consequence of researchers', reviewers' and journal editors' interpretations of research. Researchers' own interpretations of the conclusiveness of a research finding, and their belief about the likely interpretations of reviewers and editors, influence their decisions to report in detail or to attempt publication at all. Similarly, reviewers' and editors' interpretations determine their decisions about whether space will be allotted to provide detailed reporting of any particular (e.g., nonsignificant) result. Thus models of publication selection bias must be based on models of the interpretation of research results. Studies of the interpretation of research findings are therefore a promising source of ideas for models of selection bias.

Interpretative considerations such as the perceived importance of a research question or the perceived adequacy of a research paradigm undoubtedly affect the probability of selection. However, meta-analyses are unlikely to combine studies of fundamentally different research questions or use markedly different research designs. Consequently, the effects of these variables on selection are likely to be relatively constant across studies in any particular meta-analysis. Because selection effects that are unrelated to study outcomes (e.g., are constant across studies) do not introduce bias, selection effects due to these interpretative considerations do not have important implications for meta-analysis.

More important for meta-analysis is the issue of interpretation of the results of statistical analyses. Systematic research on the interpretation of statistical analyses by psychological researchers has led to several findings (e.g., Rosenthal and Gaito, 1963, 1964; Nelson, Rosenthal and Rosnow, 1986). One finding is that researchers' perceptions about the conclusiveness of research results is strongly related to the p-value. A second and rather surprising finding is that the magnitude of the estimated effect is unrelated to the perceived conclusiveness of the result. Given two results with the same p-values, the one with a larger sample size (and hence a smaller observed effect magnitude) is typically perceived to be more conclusive. A third finding is that the relationship between perceived conclusiveness of results and p-values is not smooth but is subject to "cliff effects" near conventionally used a priori levels of significance such as $\alpha = 0.05$ and $\alpha = 0.01$. In particular, there is a large change in perceived conclusiveness for a smaller change in p-value near 0.05. Thus a result with a p-value of 0.045 is perceived as much more conclusive than a result with a p-value of 0.055, but a pair of results with p-values of 0.045 and 0.035 (or 0.055 and 0.065) are perceived as about equally conclusive.

Researchers, reviewers and editors presumably select research results for potential publication in part on the basis of their perception (interpretation) of the conclusiveness of the findings. The model of interpretation and hence selection that emerges from the research findings above is one in which conclusiveness and hence the probability of selection is a function of the p-value rather than the effect size. Moreover, in this model the relation between p-value and conclusiveness is subject to jump discontinuities or "cliff effects." We introduce both of these features in the selection model given below.

Weight Function

Suppose that the observation associated with each study has a *weight function* $w(X_i)$ which determines the probability of being observed. As mentioned above, it is reasonable to posit that the weight function depends on X_i through the p-value. It is probably unreasonable to assume that much is known about the functional form of the weight function. Consequently, it is desirable to estimate the weight function using a very flexible functional form that permits the data to reveal the shape of the function. We posit that the weight function is a step function with discontinuities (steps) at points determined a priori. Suppose there are k steps. Denote the left and right endpoints of the jth step by a_{j-1} and a_j, respectively, with $a_0 = 0$ and $a_k = 1$.

Then, $w(p_i)$ is constant on the interval $a_{j-1} < p \leq a_j$. Denote the value of $w(p_i)$ on the ith interval as ω_i. We expect the weight function, as a function of p, to be the same for all studies. That is,

$$w(p_i) = \begin{cases} \omega_1, & \text{if } 0 < p_i \leq a_1, \\ \omega_j, & \text{if } a_{j-1} < p_i \leq a_j, \\ \omega_k, & \text{if } a_{k-1} < p_i \leq 1\,. \end{cases}$$

Note that in the absence of information other than just the X_i values, we can only determine the relative weights. Thus it is convenient and consistent with our model of publication bias to assign one of the ω_i values. A logical choice is to set $\omega_1 = 1.0$. This constraint implies that the ω_i values represent the chance that an estimate with a given p-value is observed *relative to the chance that studies with* $p < a_1$ is observed. Alternatively, *if* the most highly significant results are essentially always observed, the ω_i values may be interpreted as the chance that estimates with a given p-value would be observed.

In estimation, it is necessary to define the weight function as a function of the X_i. The weight function as a function of X_i depends on the study index i because the p-value depends on both X_i and σ_i. Hence

$$(3) \qquad w(X_i, \sigma_i) = \begin{cases} \omega_1, & \text{if } -\sigma_i\,\Phi^{-1}(a_1/2) < X_i \leq \infty \\ & \text{and } X_i > 0, \\ \omega_j, & \text{if } -\sigma_i\,\Phi^{-1}(a_j/2) < X_i \leq -\sigma_i\,\Phi^{-1}(a_{j-1}/2) \\ & \text{and } X_i > 0, \\ \omega_k, & \text{if } 0 < X_i \leq -\sigma_i\,\Phi^{-1}(a_{k-1}/2), \\ \omega_1, & \text{if } -\infty \leq X_i < \sigma_i\,\Phi^{-1}(a_1/2) \\ & \text{and } X_i < 0, \\ \omega_j, & \text{if } \sigma_i\,\Phi^{-1}(a_{j-1}/2) \leq X_i < \sigma_i\,\Phi^{-1}(a_j/2) \\ & \text{and } X_i < 0, \\ \omega_k, & \text{if } \sigma_i\,\Phi^{-1}(a_{k-1}/2) \leq X_i < 0\,. \end{cases}$$

The assumption that the discontinuities are known a priori, while strong, does not seem unreasonable given what is known about the social psychology of the interpretation of statistical hypothesis tests. Social conventions in research communities dictate that p-values such as 0.05, 0.01, 0.005 and 0.001 have particular salience for interpretation. Because publication bias depends on interpretation, these values are implicated as points of discontinuity in the weight function.

Likelihood

The weighted probability density of X_i given the weight function $w(X_i, \sigma_i)$ and parameters Δ, σ and $\omega = (\omega_1, \ldots, \omega_k)'$ is

$$(4) \qquad f(X_i \mid \Delta, \sigma, \omega) = \frac{w(X_i, \sigma_i)\,\phi\!\left(\dfrac{X_i - \Delta}{\eta_i}\right)}{\eta_i A_i(\Delta, \sigma_i, \omega)}\,,$$

where

$$A_i(\Delta, \sigma_i, \omega) = \int_{-\infty}^{\infty} \eta_i^{-1}\, w(X_i, \sigma_i)\,\phi\!\left(\frac{X_i - \Delta}{\eta_i}\right) dX_i\,,$$

and $n_i^2 = \sigma_i^2 + \sigma^2$. Note that A_i is the sum of normal integrals over the regions where $w(X_i, \sigma_i)$ is a constant. Thus

$$A_i(\Delta, \sigma, \omega) = \sum_{j=1}^{k} \omega_j B_{ij}(\Delta, \sigma)$$

where $B_{ij}(\Delta, \sigma)$ is the probability that a normally distributed random variable with mean Δ and variance η_i^2 is assigned weight value ω_j, that is,

$$(5) \qquad \begin{aligned} B_{i1} &= 1 - \Phi[(b_{i1} - \Delta)/\eta_i] + \Phi[(-b_{i1} - \Delta)/\eta_i]\,, \\ B_{ij} &= \Phi[(b_{i,j-1} - \Delta)/\eta_i] - \Phi[(b_{ij} - \Delta)/\eta_i] \\ &\quad + \Phi[(-b_{ij} - \Delta)/\eta_i] - \Phi[(-b_{i,j-1} - \Delta)/\eta_i]\,, \qquad 1 < j < k, \\ B_{ik} &= \Phi[(b_{i,k-1} - \Delta)/\eta_i] - \Phi[(-b_{i,k-1} - \Delta)/\eta_i]\,, \end{aligned}$$

where the b_{ij} are the left endpoints of the intervals of positive X values assigned weight ω_j in the ith study, that is,

$$b_{ij} = -\sigma_i\,\Phi^{-1}(a_j/2).$$

The joint likelihood for the data $\boldsymbol{X} = (X_1, \ldots, X_n)'$ is

$$\ell(\Delta, \sigma, \omega \mid \boldsymbol{X}) = \prod_{i=1}^{n} \frac{w(X_i, \sigma_i)\phi\!\left(\dfrac{X_i - \Delta}{\eta_i}\right)}{\eta_i A_i(\Delta, \eta_i, \omega)}\,,$$

and the log likelihood is

$$(6) \qquad \begin{aligned} L = \log(\ell) = c + \sum_{i=1}^{n} \log w_i(X_i, \omega) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i - \Delta}{\eta_i}\right)^2 \\ - \sum_{i=1}^{n} \log(\eta_i) - \sum_{i=1}^{n} \log\left[\sum_{j=1}^{k} \omega_j B_{ij}(\Delta, \sigma)\right]. \end{aligned}$$

Recalling that $\omega_1 = 1$, the likelihood equations for $\omega_2, \ldots, \omega_k$ are

$$(7) \qquad \frac{\partial L}{\partial \omega_j} = \sum_{i=1}^{n} \frac{I(i,j)}{\omega_j} - \sum_{i=1}^{n} \frac{B_{ij}}{\sum_{l=1}^{k} \omega_l B_{il}} = 0, \quad j = 2, \ldots, k,$$

where $I(i,j)$ is an indicator variable taking the value 1 if X_i is given weight ω_j and zero otherwise. Alternatively, the first sum can be written using n_j, the number of X values taking weight j which gives

$$\frac{\partial L}{\partial \omega_j} = \frac{n_j}{\omega_j} - \sum_{i=1}^{n} \frac{B_{ij}}{\sum_{l=1}^{k} \omega_l B_{il}} = 0\,.$$

The likelihood equations for Δ and σ^2 are slightly more complex because the B_{ij} depend on both Δ and σ^2. The likelihood equation for Δ is

$$(8) \qquad \frac{\partial L}{\partial \Delta} = \sum_{i=1}^{n} \frac{X_i - \Delta}{\eta_i^2} - \sum_{i=1}^{n} \frac{\sum_{j=1}^{k} \omega_j {B^{\Delta}}_{ij}}{\sum_{j=1}^{k} \omega_j B_{ij}} = 0\,,$$

where B_{ij}^{Δ} is the derivative of B_{ij} with respect to Δ, given by

$$\begin{aligned}
B_{i1}^{\Delta} &= \eta_i^{-1}\{\phi[(b_{i1}-\Delta)/\eta_i] - \phi[(-b_{i1}-\Delta)/\eta_i]\} \\
B_{ij}^{\Delta} &= \eta_i^{-1}\{\phi[(b_{ij}-\Delta)/\eta_i] + \phi[(-b_{i,j-1}-\Delta)/\eta_i] \\
&\quad -\phi[b_{i,j-1}-\Delta)/\eta_i] - \phi[(-b_{ij}-\Delta)/\eta_i]\}, \qquad 1<j<k, \\
B_{ik}^{\Delta} &= \eta_i^{-1}\{\phi[(-b_{i,k-1}-\Delta)/\eta_i] - \phi[(b_{i,k-1}-\Delta)/\eta_i]\}.
\end{aligned} \tag{9}$$

The likelihood equation for σ^2 is

$$\begin{aligned}
\frac{\partial L}{\partial \sigma^2} &= \frac{1}{2}\sum_{i=1}^{n}\frac{(X_i-\Delta)^2}{\eta_i^4} - \sum_{i=1}^{n}\frac{1}{2\eta_i^2} - \sum_{i=1}^{n}\frac{\Sigma_{j=1}^{k}\omega_j B_{ij}^{\sigma}}{\Sigma_{j=1}^{k}\omega_j B_{ij}} \\
&= 0,
\end{aligned} \tag{10}$$

where the B_{ij}^{σ} are the partial derivatives of B_{ij} with respect to σ^2 given by

$$\begin{aligned}
B_{i1}^{\sigma} &= \frac{1}{2}\eta_i^{-2}[c_{i1}\phi(c_{i1}) - d_{i1}\phi(d_{i1})], \\
B_{ij}^{\sigma} &= \frac{1}{2}\eta_i^{-2}[c_{ij}\phi(c_{ij}) + d_{i,j-1}\phi(d_{i,j-1}) \\
&\quad - c_{i,j-1}\phi(c_{i,j-1}) - d_{ij}\phi(d_{ij})], \qquad 1<j<k, \\
B_{ik}^{\sigma} &= \frac{1}{2}\eta_i^{-2}[d_{i,k-1}\phi(d_{i,k-1}) - c_{i,k-1}\phi(c_{i,k-1})],
\end{aligned} \tag{11}$$

where

$$c_{ij} = (b_{ij}-\Delta)/\eta_i \tag{12}$$

and

$$d_{ij} = (-b_{ij}-\Delta)/\eta_i. \tag{13}$$

The second derivatives of the likelihood, useful for computing solutions to the likelihood equations and the information matrix, are only slightly more complicated. They are given in the Appendix.

Estimation

One approach to parameter estimation is the simultaneous estimation of Δ, σ and ω via the Newton-Raphson method. This involves the inversion of a $(k+1) \times (k+1)$ Hessian matrix at each iteration, but is quite feasible as long as $(k-1)$, the number of steps in the weight function, is not too large.

An alternative approach to estimation of Δ, σ and ω is the use of the EM algorithm. The first (E) step is as follows. Given initial values of $\omega^{(0)}$, compute estimates (the expectation $\Delta^{(1)}$ and $\sigma^{(1)}$) of Δ and σ given $\omega^{(0)}$ and the data. The first (M) step involves computing an estimate $\omega^{(0)}$ of ω given the estimates $\Delta^{(0)}$ and $\sigma^{(0)}$ computed in the E-step. Subsequent E and M steps of the algorithm follow the same pattern. The jth E-step consists of computing estimates $\Delta^{(j)}$ and $\sigma^{(j)}$ given the estimate $\omega^{(j-1)}$ from the previous step. The jth M-step then consists of computing the estimate $\omega^{(j)}$ from $\Delta^{(j)}$, $\sigma^{(j)}$ and the data. This EM algorithm requires numerical computation of the estimates in each step, but the computations are less demanding than direct estimation of all parameters because the $(k-1) \times (k-1)$ Hessian matrix to be inverted at each step is smaller than that required for direct estimation. However, convergence has been very slow in practice, which effectively cancels the computational advantages of this method since many more iterations are needed.

Graphical Procedures for Exploring the Weight Function

One way to explore the likelihood of publication bias is to examine the observed distribution of p-values and to compare that distribution with the distribution that would be expected given no publication bias. Such a comparison also provides insight about the likely shape of the weight function. For example, regions where the observed and expected frequency of p-values differ substantially correspond to regions where the value of the weight function departs from one. Such comparisons can also be the basis for testing the hypothesis that there is no publication bias (i.e., that $\omega_1 = \cdots = \omega_k = 1$) *assuming that the random effects specification of the model is correct.*

The probability density function of the p-value from study i depends on the distribution of the test statistic z under null and alternative hypotheses (Pearson, 1938). In particular, the probability function of the p from study i is given by

$$f_i(p|\sigma_i, \Delta, \sigma) = \frac{\sigma_i\,\phi\{[\mathrm{sign}(X_i)\sigma_i\,\Phi^{-1}(p/2) + \Delta]/\eta_i\}}{\eta_i\,\phi\{[\mathrm{sign}(X_i)\,\sigma_i\,\Phi^{-1}(p/2)]/\sigma_i\}},$$

where sign(X_i) takes the value +1 if X_i is nonnegative and −1 if X_i is negative. A p-value picked at random from $p_1, \ldots, p_n$ has a probability density that is a mixture of the probability densities of $p_1, \ldots, p_n$, namely

$$f(p|\sigma_1, \ldots, \sigma_n, \Delta, \sigma) = \frac{1}{n}\sum_{i=1}^{n} f_i(p|\sigma_i, \Delta, \sigma). \tag{14}$$

Expression 14 gives the theoretical probability density for the p-values that would be expected if there is no publication bias. It provides a reference density against which to compare the empirical frequency distribution of observed values, and it can be integrated numerically to yield the expected proportion of responses in any particular interval of p-values.

The distribution (14) of the p-values depends on Δ and σ, the mean and variance of the random effects, which are unknown. To use these results in practice, it is therefore necessary to estimate Δ and σ and to substitute the estimates for the parameter values in (14). Simple, consistent estimates of Δ and σ^2 (Hedges and Olkin, 1985) are

$$\hat{\sigma}_0^2 = \sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{n-1} - \sum_{i=1}^{n}\frac{\sigma_i^2}{n} \tag{15}$$

and

$$\hat{\Delta}_0 = \frac{\Sigma_{i=1}^{n} X_i/(\hat{\sigma}_0^2 + \sigma_i^2)}{\Sigma_{i=1}^{n} 1/(\hat{\sigma}_0^2 + \sigma_i^2)}, \tag{16}$$

where $\overline{X}$ is the unweighted mean of $X_1, \ldots, X_n$. Graphical procedures may be used for comparing the empirical and the expected density $f(p|\sigma_1, \ldots, \sigma_n, \hat{\Delta}_0, \hat{\sigma}_0)$ or the empirical and expected cumulative distribution function. For example, plotting the observed versus the expected density or cumulative distribution functions provides a way of determining ranges of p-values that occur more frequently than would be expected if there were no publication bias.

Testing for Publication Bias

The comparison of observed and expected distributions of p-values provides the basis for conditional tests of publication bias. We emphasize that the tests are *conditional* in that they depend on the assumption that the random effects are normally distributed and that the true weight function can be adequately approximated by a step function with the steps that were specified a priori. The specification of the distribution of the random effects is critical because it affects the computation of the *expected* distribution of the p-values in the absence of publication bias. If the random effects have a nonnormal distribution, the tests would not be valid.

Test Based on Grouped Frequencies. One test is simply a chi-squared test of goodness of fit of the observed p-values to the expected p-value distribution. Given k intervals defined by the cutpoints $0 \equiv a_0 < a_1 < \cdots < a_k \equiv 1$, tabulate the observed number O_j of p-values in the jth interval $[a_{j-1}, a_j]$ and compute the expected number of p-values in the jth interval via

$$(17) \qquad \begin{aligned} E_j &= n \int_{a_{j-1}}^{a_j} f(p|\sigma_1, \ldots, \sigma_n, \hat{\Delta}_0, \hat{\sigma}_0)dp \\ &= \sum_{i=1}^{n} B_{ij'}(\hat{\Delta}_0, \hat{\sigma}_0), \end{aligned}$$

where $B_{ij}(\Delta, \sigma)$ is given by (5).

Then compute the chi-squared goodness of fit statistic X^2 via

$$(18) \qquad X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}.$$

If $\omega_2 = \cdots = \omega_k = 1$ (that is, if there is no publication bias) and if the number of studies n is large compared with k, then X^2 will have approximately a chi-squared distribution with $k - 1$ degrees of freedom. We reject the hypothesis of no publication bias at significant level α if X^2 exceeds the $100(1 - \alpha)$ percent critical value of the chi-squared distribution.

Note that the ratio O_j/E_j for the jth interval of p-values is a crude estimate of the weight ω_j for the interval. Intervals in which O_j/E_j departs substantially from 1 would be identified as p-value intervals in which selection effects are evident.

Test Based on the Likelihood Ratio Criterion. The conditional test described above is easy to implement and is a natural extension of graphical procedures to compare observed and expected distributions of p-values. Another, more elegant, conditional test is based on the likelihood ratio criterion. This test uses the likelihood ratio test statistic

$$(19) \qquad G^2 = 2[L(\hat{\Delta}, \hat{\sigma}, \hat{\omega}) - L(\hat{\Delta}_0, \hat{\sigma}_0, 1)]$$

where $L(\Delta, \sigma, \omega)$ is the log likelihood (6), $\hat{\Delta}$, $\hat{\sigma}$ and $\hat{\omega}$ are the maximum likelihood estimates of Δ, σ and ω with no constraints, $\hat{\Delta}_0$ and $\hat{\sigma}_0$ are consistent (e.g., maximum likelihood) estimates of Δ and σ under the assumption of no publication bias and 1 is a vector of $k - 1$ ones. If

$$H_0: \omega_2 = \cdots = \omega_k = 1$$

is true (that is, if there is no publication bias), G^2 has approximately a chi-squared distribution with $k - 1$ degrees of freedom. We reject H_0 at significance level α if G exceeds the $100(1 - \alpha)$ percentage point of the chi-square distribution with $k - 1$ degrees of freedom.

The likelihood ratio test statistic is easy to compute in conjunction with computation of the likelihood estimates of Δ, σ and ω and the estimates $\hat{\Delta}_0$ and $\hat{\sigma}_0$ given in (15) and (16) of Δ and σ assuming no publication bias.

APPLYING THE RESULTS OF THIS PAPER

Indexes of Effect Magnitude

Research synthesis in the social and behavioral sciences typically have used either the standardized mean difference (Glass' effect size) or the product-moment correlation as the index of effect magnitude. The results of each study are typically expressed as estimates of one of these parameters and combined across studies in the meta-analysis. The results in this paper can be applied to either of these indexes of effect magnitude because estimates of both correlation coefficients and standardized mean differences, after suitable transformations, are approximately normally distributed with known variance.

Correlation Coefficient. Suppose that the m_i independent observations $(X_{i1}, Y_{i1}), \ldots, (X_{im_i}, Y_{im_i})$ within the ith study are sampled from a bivariate normal distribution with unknown correlation coefficient ρ_i. Then the sample product-moment correlation coefficient r_i is an estimate of ρ_i and the Fisher z-transform of r_i, given by

$$z_i = z(r_i) = 0.5 \log\left[\frac{1 + r_i}{1 - r_i}\right],$$

is approximately normally distributed about

$$\zeta_i = z(\rho_i) = 0.5 \log\left[\frac{1 + \rho_i}{1 - \rho_i}\right]$$

with variance $1/(m_i - 3)$. The results of this paper are applied in meta-analyses using the correlation coefficient as the index of effect magnitude by defining

$$X_i \equiv z_i, \qquad \sigma_i^2 \equiv \frac{1}{m_i - 3}$$

and letting

$$\zeta_i \sim N(\Delta, \sigma^2)\,.$$

Standardized Mean Difference. Suppose that the m_i independent observations in the experimental and control groups of the ith study are $Y_{i1}^E, \ldots, Y_{im_i}^E$ and $Y_{i1}^C, \ldots, Y_{im_i}^C$, respectively. Then if

$$Y_{ij}^E \sim N(\mu_i^E, \sigma_i^2), \qquad Y_{ij}^C \sim N(\mu_i^C, \sigma_i^2), \qquad j = 1, \ldots, m_i\,,$$

the population effect size δ_i and its sample estimate d_i are given by

$$\delta_i = \frac{\mu_i^E - \mu_i^C}{\sigma_i}, \qquad d_i = \frac{\overline{Y}_i^E - \overline{Y}_i^C}{s_i},$$

where $\overline{Y}_i^E$, $\overline{Y}_i^C$ and s_i^2 are the experimental and control group sample means and the sample estimate of the within-group variance, respectively. If m_i is not too small (e.g., $m_i \geq 10$), then d_i is approximately normally distributed about δ_i with variance

$$\frac{2}{m_i}\left(1 + \frac{d_i^2}{4}\right).$$

If the effect size is small, then d_i will have little effect on the variance and the variance can be treated as essentially known. In this case, the results of this paper can be applied by defining

$$X_i \equiv d_i, \qquad \sigma_i^2 \equiv \frac{2}{m_i}\left(1 + \frac{d_i^2}{4}\right)$$

and letting

$$\delta_i \sim N(\Delta, \sigma^2)\,.$$

Alternatively, a more elegant approach is to use the variance stabilizing transformation given by Hedges and Olkin (1985):

$$T_i = 2 \sinh^{-1} (d_i/2\sqrt{2}), \qquad \theta_i = 2 \sinh^{-1} (\delta_i/2\sqrt{2})\,.$$

Then T_i is approximately normally distributed with a mean of θ_i and a variance of $1/m_i$. The results of this paper can then be applied by defining

$$\mathrm{X}_i \equiv T_i, \qquad \sigma_i^2 \equiv 1/m_i$$

and letting

$$\theta_i \sim N(\Delta, \sigma^2)\,.$$

EXAMPLE

The methods described in this paper were applied to the data from a meta-analysis of 755 studies of the validity of the General Aptitude Tests Battery (GATB), a cognitive test measuring several mental abilities. These studies were carried out by the United States Employment Service (USES) to determine whether the GATB was a useful predictor of job performance in a wide variety of employment settings. Analyses of this data set have been the basis for the policy decision to use the GATB for referral of job applicants for job placement in the USES offices nationally. That decision proved controversial and ultimately led to a lawsuit against USES on the grounds that the data from the studies were insufficiently conclusive to warrant the national policy. A National Research Council panel was convened to study the problem (Hartigan and Wigdor, 1989). One of the questions that emerged was whether there was a systematic tendency to selectively report GATB validity studies.

The data reported for each study consists of a correlation coefficient between a GATB ability scale and a measure of job performance. Although the GATB has nine ability scales, only results for the general ability scale are examined here. (General ability and composite ability scales that rely heavily on it figure most prominently in the controversy over the validity of the GATB.) The analysis was performed on the Fisher z-transformed correlations.

Assuming no selection, the estimate of the average (z-transformed) correlation Δ would be $\hat{\Delta}_0 = 0.26$ and the estimated variance component of the population correlations would be $\hat{\sigma}_0^2 = 0.012$ corresponding to $\hat{\sigma}_0 = 0.11$. We implemented the selection model by using 10 p-value intervals defined by $a_0 = 0.0, a_1 = 0.001$, $a_2 = 0.005$, $a_3 = 0.01$, $a_4 = 0.02$, $a_5 = 0.05$, $a_6 = 0.01$, $a_7 = 0.20, a_8 = 0.30, a_9 = 0.50$ and $a_{10} = 1.0$. We began the analysis by comparing the observed and the expected number of reports in each interval. Table 1 presents the number of observed p-values falling into each interval and the number expected if there were no publication bias. Note that the observed number corresponds closely to the number expected in each interval but that of the largest p-values $0.5 \leq p < 1.0$, where only about 70% of the expected number of p-

TABLE 1
Number of observed p-values falling into each interval and the number expected if there were no selection bias

	Observed		Expected		
Interval	Number	%	Number	%	Chi-square
0.0–0.001	179	23.7	187.4	24.8	0.38
0.001–0.005	80	10.6	75.0	9.9	0.34
0.005–0.01	43	5.7	41.6	5.5	0.05
0.01–0.02	44	5.8	48.0	6.4	0.33
0.02–0.05	81	10.7	73.9	9.8	0.69
0.05–0.10	80	10.6	63.8	8.5	4.14
0.10–0.20	68	9.0	70.0	9.3	0.06
0.20–0.30	45	6.0	43.7	5.8	0.04
0.30–0.50	61	8.1	58.3	7.7	0.13
0.50–1.0	74	9.8	93.4	12.4	4.04

Note: The total chi-square is 10.19 with nine degrees of freedom, $p = 0.25$.

TABLE 2
Maximum likelihood estimates of the parameters under the selection model

Interval	$\hat{\omega}_j$	SE($\hat{\omega}_j$)	O_j/E_j
0.0–0.001	1.0	–	–
0.001–0.005	1.03	0.16	1.10
0.005–0.01	0.98	0.19	1.03
0.01–0.02	0.86	0.17	0.92
0.02–0.05	1.00	0.18	1.09
0.05–0.10	1.13	0.22	1.25
0.10–0.20	0.86	0.19	0.97
0.20–0.30	0.91	0.22	1.04
0.30–0.50	0.91	0.22	1.05
0.50–1.0	0.68	0.18	0.79

Note: In this analysis, the maximum likelihood estimates of Δ and σ^2 are $\hat{\Delta} = 0.25$, $\hat{\sigma}^2 = 0.011$ with standard errors computed from the information matrix as SE($\hat{\Delta}$) = 0.012 and SE($\hat{\sigma}^2$) = 0.0013.

values are observed. However, computing the overall chi-square test we see that the observed pattern of p-values does not differ significantly from that expected if there were no publication bias (given the random effects model and the choice of p-value intervals). Computing the statistics (18) and (19), we obtain the values of $X^2 = 10.19$ and $G^2 = 10.17$, respectively. Comparing these values with the percentage points of the chi-square distribution with nine degrees of freedom, we see that values this large would arise due to chance more than 25% of the time.

Computing the maximum likelihood estimate of ω, Δ and σ^2 under the selection model, we obtain

$$\hat{\Delta} = 0.25, \hat{\sigma}^2 = 0.011 \text{ and } \hat{\sigma} = 0.11$$

which do not differ appreciably from the values obtained under the model, assuming no selection bias. The maximum likelihood estimates of the weights are given in Table 2 along with their standard errors obtained as the square root of the diagonal elements of the inverse of the information matrix. All but one of the weights are estimated as near 1.0. Thus the selection model would not be expected to greatly influence the estimates of Δ and σ^2. It is interesting that the weights are so poorly estimated. Each has an estimated standard error of approximately 0.2, which suggests that the data contain relatively little information about selection. It is also interesting that the crude estimates of the weights given by O_j/E_j (at least in this example) are quite close to the maximum likelihood estimates of the weights.

This analysis suggests that there is little evidence that substantial publication selection bias exists in these data. The chi-square test suggested only very meager evidence of selection bias. Perhaps more important is the fact that the estimates of Δ and σ^2 under the selection model were essentially identical to those obtained from an estimation procedure that assumed no selection.

CONCLUSIONS

The methods described in this paper provide one way of examining a set of data to determine if it is consistent with the pattern of observations that might exist when publication selection is operating. While the methods provide a model-based correction for the effects of publication bias on estimates of effect size, such estimates should be interpreted cautiously when there is evidence of substantial selection. The validity of any model-based correction depends on the accuracy of the model. While the model proposed seems generally sensible it may not be accurate in detail. Thus these methods are probably best used to provide a broad indication of whether selection is operating and, if so, what its gross effects have been on estimation.

APPENDIX

The second partial derivatives of the likelihood (6) for computing the information matrix are

$$\frac{\partial^2 L}{\partial \omega_j \, \partial \omega_\ell} = -\,\delta_\ell^j \sum_{i=1}^{n} \frac{I(i,j)}{\omega_j^2} + \sum_{i=1}^{n} \frac{B_{ij} B_{i\ell}}{[\Sigma_{m=1}^{k} \omega_m B_{im}]^2},$$

where

$$\delta_\ell^j = \begin{cases} 1, & \text{if } j = \ell, \\ 0, & \text{if } j \neq \ell, \end{cases}$$

$$\frac{\partial^2 L}{\partial \omega_j \, \partial \Delta} = \sum_{i=1}^{n} \frac{(\Sigma_{m=1}^{k} \omega_m B_{im}) B_{ij}^{\Delta} - B_{ij} (\Sigma_{m=1}^{k} \omega_m B_{im}^{\Delta})}{(\Sigma_{m=1}^{k} \omega_m B_{im})^2},$$

$$\frac{\partial^2 L}{\partial \omega_j \, \partial \sigma^2} = \sum_{i=1}^{n} \frac{(\Sigma_{m=1}^{k} \omega_m B_{im}) B_{ij}^{\sigma} - B_{ij} (\Sigma_{m=1}^{k} \omega_m B_{im}^{\sigma})}{(\Sigma_{m=1}^{k} \omega_m B_{im})^2},$$

$$\frac{\partial^2 L}{\partial \Delta^2} = -\sum_{i=1}^{n} \frac{1}{\eta_i^2} - \sum_{i=1}^{n} \frac{(\Sigma_{m=1}^{k} \omega_m B_{im})(\Sigma_{m=1}^{k} \omega_m B_{im}^{\Delta\Delta}) - (\Sigma_{m=1}^{k} \omega_m B_{im}^{\Delta})^2}{(\Sigma_{m=1}^{k} \omega_m B_{im})^2},$$

$$\frac{\partial^2 L}{\partial \Delta \partial \sigma^2} = -\sum_{i=1}^{n} \frac{X_i - \Delta}{\eta_i^4} - \frac{(\Sigma_{m=1}^{k} \omega_m B_{im})(\Sigma_{m=1}^{k} \omega_m B_{im}^{\Delta\sigma})}{(\Sigma_{m=1}^{k} \omega_m B_{im})^2} + \frac{(\Sigma_{m=1}^{k} \omega_m B_{im}^{\sigma})(\Sigma_{m=1}^{k} \omega_m B_{im}^{\Delta})}{(\Sigma_{m=1}^{k} \omega_m B_{im})^2},$$

$$\frac{\partial^2 L}{\partial (\sigma^2)^2} = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\eta_i^4} - \sum_{i=1}^{n} \frac{(X_i - \Delta)^2}{\eta_i^6} - \sum_{i=1}^{n} \frac{(\Sigma_{m=1}^{k} \omega_m B_{im})(\Sigma_{m=1}^{k} \omega_m B_{im}^{\sigma\sigma}) - (\Sigma_{m=1}^{k} \omega_m B_{im}^{\sigma})^2}{(\Sigma_{m=1}^{k} \omega_m B_{im})^2},$$

where

$$B_{ij}^{\Delta\sigma} = -\eta_i^{-1} B_{ij}^{\sigma} - \frac{1}{2} \eta_i^{-2} B_{ij}^{\Delta},$$

$$B_{ij}^{\Delta\Delta} = 2B_{ij}^{\sigma},$$

and $B_{ij}^{\sigma\sigma}$ is given by

$$B_{i1}^{\sigma\sigma} = -\frac{3}{2}\eta_i^{-2} B_{i1}^{\sigma} - \frac{1}{4}\eta_i^{-4}[d_{i1}^3\,\phi(d_{i1}) - c_{i1}^3\,\phi(c_{i1})],$$

$$B_{ij}^{\sigma\sigma} = -\frac{3}{2}\eta_i^{-2} B_{ij}^{\sigma} - \frac{1}{4}\eta_i^{-4}[c_{i,j-1}^3\,\phi(c_{i,j-1}) - c_{ij}^3\,\phi(c_{ij}) + d_{ij}^3\,\phi(d_{ij}) - d_{i,j-1}^3\,\phi(d_{i,j-1})], \qquad 1<j<k,$$

$$B_{ik}^{\sigma\sigma} = \frac{3}{2}\eta_i^{-2} B_{ik}^{\sigma}[c_{i,k-1}^3\,\phi(c_{i,k-1}) - d_{i,k-1}^3\,\phi(d_{i,k-1})].$$

ACKNOWLEDGMENTS

I thank Nancy Clements, Michael Linacre and Jack Vevea for their help in programming and data analysis. A computer program for carrying out the estimation described in this paper is available from the author.

REFERENCES

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin* **66** 432–437.

Begg, C. B. and Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *J. Roy. Statist. Soc. Ser. A* **151** 1–27.

Berlin, J. A., Begg, C. B. and Louis, T. A. (1989). An assessment of publication bias using a sample of published clinical trials. *J. Amer. Statist. Assoc.* **84** 381–392.

Bozarth, J. D. and Roberts, R. R. (1972). Signifying significant significance. *American Psychologist* **27** 774–775.

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

Chase, L. J. and Chase, R. B. (1976). Statistical power analysis of applied psychological research. *Journal of Applied Psychology* **61** 234–237.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology* **65** 145–153.

Cooper, H. M. (1984). *The Integrative Research Review.* Sage, Beverly Hills.

Coursol, A. and Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology* **17** 136–137.

Dawes, R. M., Landman, J. and Williams, M. (1984). Discussion on meta-analysis and selective publication bias. *American Psychologist* **39** 75–78.

Dear, K. B. G. and Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statist. Sci.* **7** 237–245.

Devine, E. C. and Cook, T. D. (1983). A meta-analysis of effects of psychoeducational interventions on length of postsurgical hospital stay. *Nursing Research* **32** 267–274.

Dickerson, K., Chan, S., Chalmers, T. C., Sacks, H. S. and Smith, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials* **8** 343–353.

Eagly, A. H. and Carli, L. L. (1981). Sex of researchers and sex typed communications as determinants of influenceability: A meta-analysis of social influence studies. *Psychological Bulletin* **90** 1–20.

Glass, G. V., McGaw, B. and Smith, M. L. (1981). *Meta-Analysis in Social Research.* Sage, Beverly Hills.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin* **82** 1–20.

Hartigan, J. A. and Wigdor, A. K. (1989). *Fairness in Employment Testing: Validity Generalization, Minority Issues and the General Aptitude Test Battery.* National Academy Press, Washington, D.C.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* **9** 61–85.

Hedges, L. V. (1989). Estimating the normal mean and variance under a publication selection model. In *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin* (L. J. Gleser, M. D. Perlman, S. J. Press and A. R. Sampson, eds.) 447–458. Springer, New York.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Academic, New York.

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statist. Sci.* **3** 109–135.

Lane, D. M. and Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British J. Math. Statist. Psych.* **31** 107–112.

Light, R. J. and Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing Research.* Harvard Univ. Press.

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology* **64** 553–557.

Nelson, N., Rosenthal, R. and Rosnow, R. L. (1986). Interpretation of significance levels by psychological researchers. *American Psychologist* **41** 1299–1301.

Orwin, R. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics* **8** 157–159.

Pearson, E. S. (1938). The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika* **30** 134–148.

Peto, R. (1987). Why do we need systematic overviews of randomized trials? *Statistics in Medicine* **6** 233–240.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyā Ser. A* **27** 311–324.

Rao, C. R. (1985). Weighted distributions arising out of methods of ascertainment: What population does a sample represent? In *A Celebration of Statistics: The ISI Centenary Volume* (A. C. Atkinson and S. E. Fienberg, eds.) 543–569. Springer, New York.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin* **86** 638–641.

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research.* Sage, Beverly Hills.

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

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

Sidman, M. (1960). *Tactics of Scientific Research.* Basic Books, New York.

Simes, R. J. (1986a). Confronting publication bias: A cohort design for meta-analysis. *Statistics in Medicine* **6** 11–30.

Simes, R. J. (1986b). Publication bias: The case for an international registry of clinical trials. *Journals of Clinical Oncology* **4** 1529–1541.

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

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

Strube, M. J. (1981). Meta-analysis and cross-cultural comparison: Sex differences in child competitiveness. *Journal of Cross Cultural Psychology* **12** 3–20.

White, K. R. (1982). The relation between socioeconomic status and achievement. *Psychological Bulletin* **91** 461–481.