PSYCHOMETRIKA—VOL. 60, NO. 3, 419–435
SEPTEMBER 1995

A GENERAL LINEAR MODEL FOR ESTIMATING EFFECT SIZE IN THE PRESENCE OF PUBLICATION BIAS

JACK L. VEVEA AND LARRY V. HEDGES

UNIVERSITY OF CHICAGO

When the process of publication favors studies with small p-values, and hence large effect estimates, combined estimates from many studies may be biased. This paper describes a model for estimation of effect size when there is selection based on one-tailed p-values. The model employs the method of maximum likelihood in the context of a mixed (fixed and random) effects general linear model for effect sizes. It offers a test for the presence of publication bias, and corrected estimates of the parameters of the linear model for effect magnitude. The model is illustrated using a well-known data set on the benefits of psychotherapy.

Key words: meta-analysis, research synthesis, publication bias, effect size, mixed models, selection models.

Introduction

There is growing agreement that the problem of research synthesis—the combining of information across replicated research studies—is of fundamental importance in a variety of scientific domains (National Research Council, 1992). Procedures for such synthesis that combine quantitative estimates of treatment effects across published studies have become known as *meta-analysis*; a large collection of such techniques has emerged (see, e.g., Hedges & Olkin, 1985, Cooper & Hedges, 1994). An assumption underlying most of those techniques is that the estimates of treatment effects available to the meta-analyst constitute a representative sample of such estimates from all research undertaken on an issue. It has long been known, however, that this assumption is often highly questionable when effect estimates are derived from the published literature. Evidence against the general validity of the assumption has taken a number of forms. Researchers have followed studies approved by institutional review boards or granting agencies and have found that failure to achieve statistical significance is strongly associated with failure to publish results (see, e.g., Dickersin, Min, & Meinert, 1991; Dickersin, Min & Meinert, 1992; Easterbrook, Berlin, Gopalan, & Matthews, 1991). Others have found that statistical significance is often a formal (Melton, 1962) or informal (Coursol & Wagner, 1986; Greenwald, 1975) criterion when editors and reviewers consider articles for publication. Studies of published literature in the social sciences tend to find surprisingly high percentages of tests rejecting at the $\alpha = .05$ level, given typical power levels of such research. (See, e.g., Bozarth & Roberts, 1972; Sterling, 1959). Finally, studies that compare effect sizes in published and unpublished literature (e.g., Dawes, Landman, & Williams, 1984; Smith, 1980; White, 1982) have found that the latter effects tend to be smaller. Thus, there is a considerable body of literature suggesting that this fundamental assumption of meta-analysis is often not strictly met. (See Begg, 1994, for an overview of the problem and a discussion of the mechanisms by which differential selection of studies with highly significant effect estimates leads to bias.)

Various solutions to this problem have been proposed. These fall into three broad

Authors' note: The contributions of the authors are considered equal, and the order of authorship was chosen to be reverse-alphabetical. Requests for reprints and for the computer program employed in this work may be directed to Jack L. Vevea, University of North Carolina at Chapel Hill, NC 27599-3270.

0033-3123/95/0900-94207$00.75/0

classes: (a) Methods that *eliminate* differential selection; (b) techniques that attempt to *detect* the presence of publication bias, so that results may be interpreted with appropriate caution; and (c) approaches that attempt to *compensate* for publication bias by establishing what the combined effect estimates would have been if censorship had not occurred. Examples of the first type include changing editorial policies through education to make editors and reviewers aware of the problem, providing avenues for the publication of studies with non-significant or negative outcomes, and developing registries of studies that have been undertaken (see, e.g., Begg & Berlin, 1988). Such approaches would clearly be most effective, but, given the realities of research in the social sciences, they are unlikely to be feasible.

Techniques for detecting publication bias have tended to be graphical. So-called "funnel plots" with the individual studies' estimates of effect magnitude on one axis, and sample size on the other were advanced by Light and Pillemer (1984). In the absence of selection, such plots should be symmetrical and funnel-shaped. When studies with large one-tailed p-values (e.g., nonsignificant studies) tend to be censored, the plot becomes skewed (the one-tailed selection pattern). If the actual underlying effect is small, and selection is based on *two-tailed* p-values, the plot may show symmetric tails, with a sparseness of studies that have small sample sizes and effects near zero (often near the center of the range of effect magnitudes). The funnel plot has also inspired a class of statistical tests based on the idea that selection implies an association between sample size (or, equivalently, sampling variability) and effect magnitude (see Begg, 1994, for examples). Figure 1 shows funnel plots of simulated data generated from identical parameters with and without strong one-tailed selection; control lines have been added that should include roughly 90 percent of the effect estimates. Although the first plot shows a sparseness of studies with negative effects (there are only two studies below the lower control line), such results can be difficult to detect when the pattern of selection is less extreme.

In practice, we have found that it is often more informative to plot effect magnitude against the conditional sampling variance rather than sample size. Sampling variance is roughly proportional to the reciprocal of sample size. This modified style of funnel plot has the effect of expanding the axis in the range where studies have small to moderate sample sizes. Since those are often the studies most likely to show selection effects, the modified plot can provide a clearer picture of publication bias. Figure 2 shows modified plots for the same data sets that were shown in Figure 1; again, approximate 90 percent control lines help expose the symmetry of the unbiased data set and the skew of the biased data. The plot of the unbiased data shows a symmetrical $\supset$-shape, with the $\supset$ centered on the mean population effect. The sparseness of small and negative effects in the biased data is more apparent now: what showed as a subtle lack of symmetry in the conventional funnel plot is much more obvious.

Methods that compensate for publication bias work by employing some statistical model for the observed effect sizes that incorporates the selection process. Such a model will comprise two parts: a model for the distribution of effect size estimates before selection occurred, and a model for the selection process that describes how that process has modified the effect distribution. Iyengar and Greenhouse (1988) proposed modeling the selection process by using weighted distributions, in which a weight function describes the likelihood of effects in a given range being observed if they occur. Hedges (1984) and Lane and Dunlap (1978) studied the effects of observing *only* studies whose mean differences were significant at $\alpha = .05$ or better. In retrospect, these cases might be considered extreme cases of weighted distributions, in which the weight is zero for a nonsignificant study, or one otherwise. Recently, attention has turned to estimating more sophisticated weight functions. Dear and Begg (1992) esti-

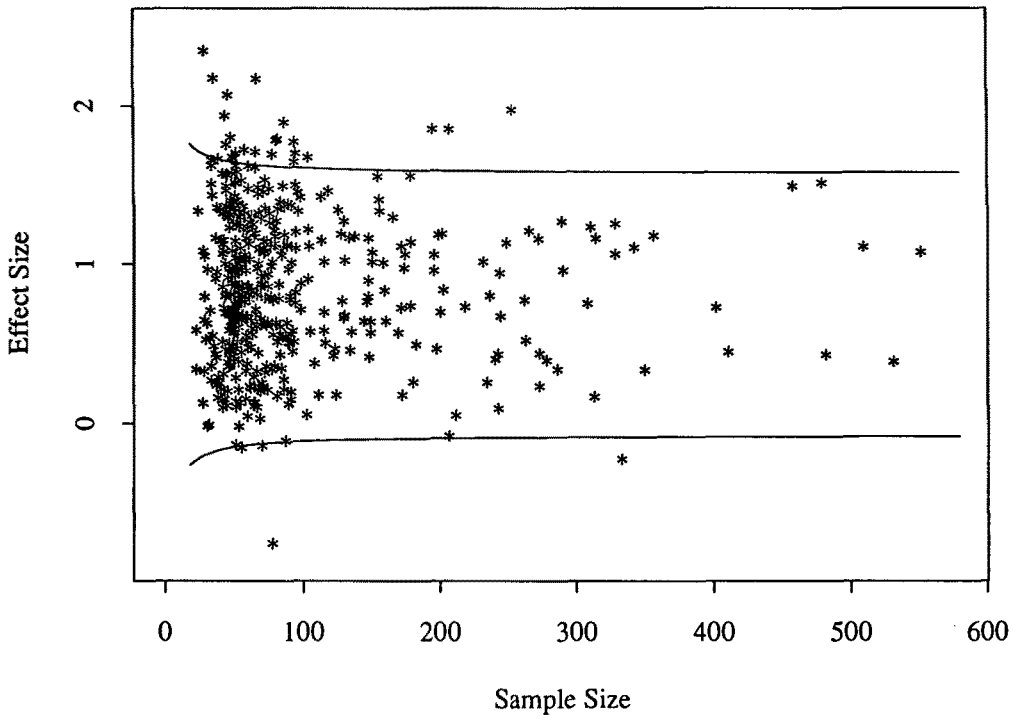

FIGURE 1A.
Funnel plots showing date sets with one-tailed selection.

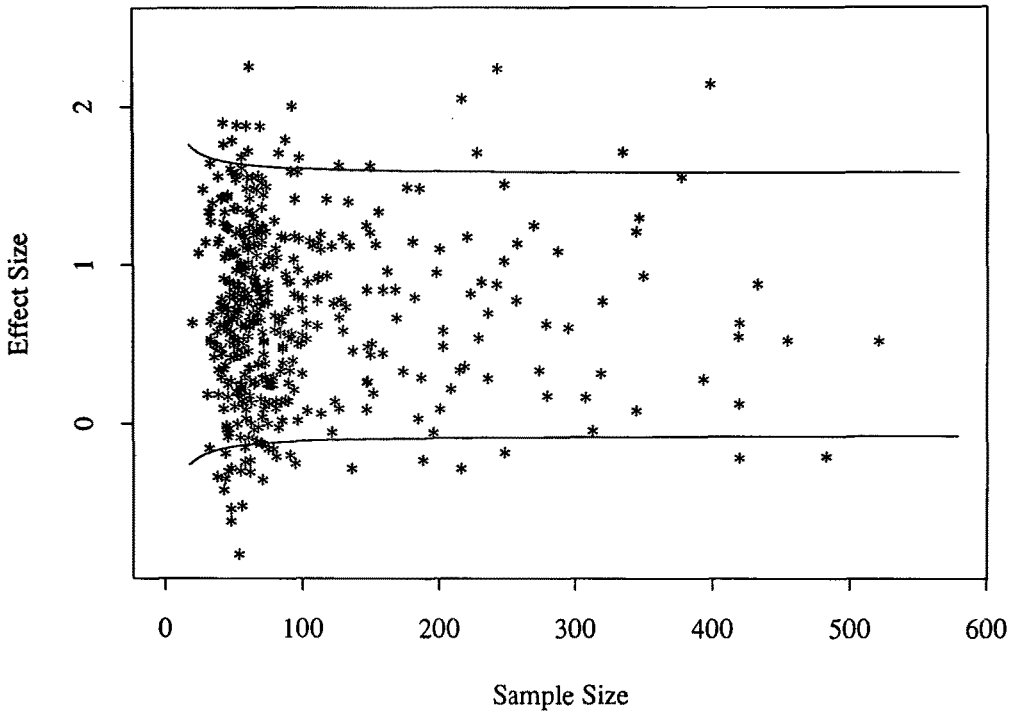

FIGURE 1B.
Funnel plots showing date sets with no selection.

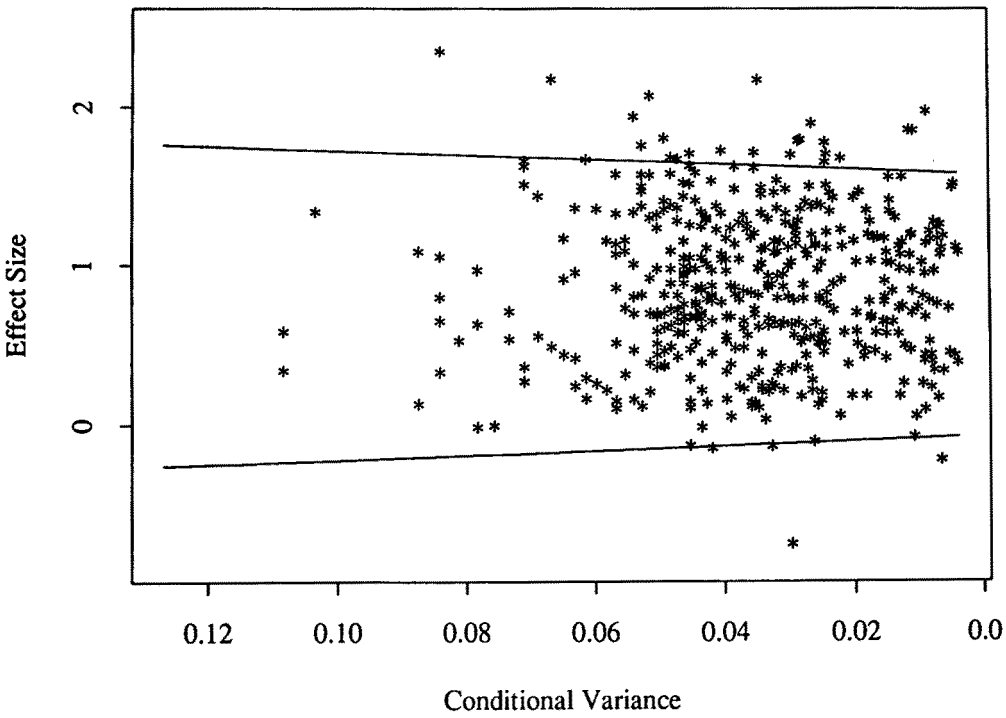

FIGURE 2A.
Modified funnel plots showing data sets with one-tailed selection.

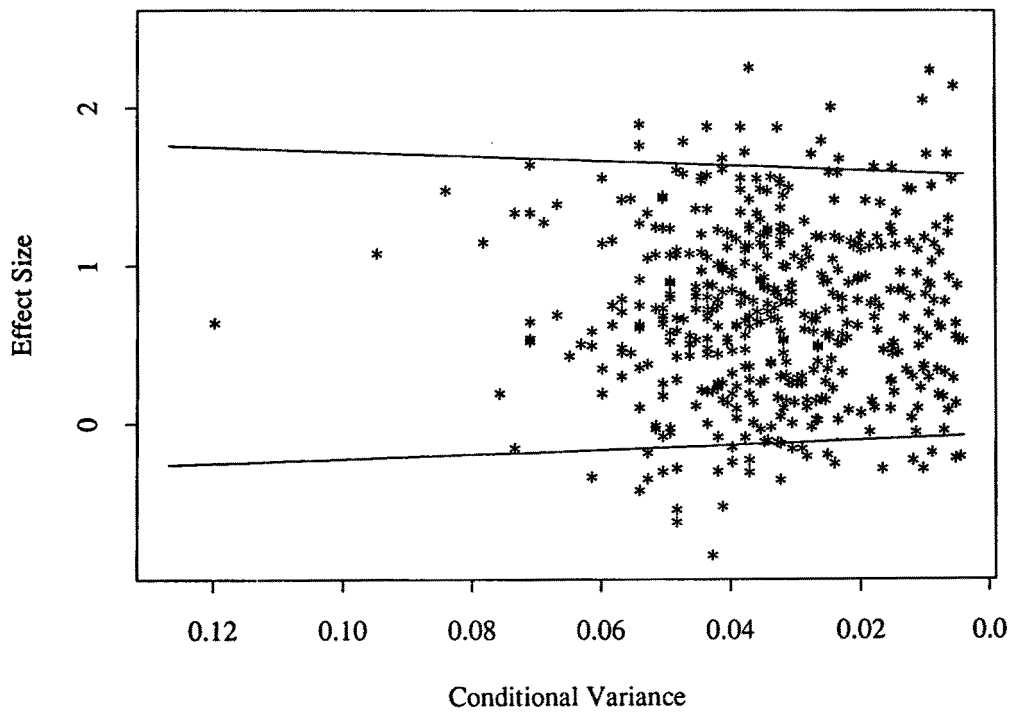

FIGURE 2B.
Modified funnel plots showing data sets with no selection.

mated a step function, with both the weights and the locations of point discontinuities determined by the data. Hedges (1992) presented a similar model, but with the discontinuities assumed to be located a points where psychologically important p-values occur (e.g., .001, .01, .05, etc.), so that only the weights need to be estimated (rather than both the weights and the points of discontinuity). In the original formulation of that model, two-tailed p-values were employed. Subsequent formulations of the model (Hedges & Vevea, 1993; Vevea, Clements & Hedges, 1993) have employed cutpoints determined by one-tailed p-values.

Model and Notation

The statistical model can be divided into two parts. The effect size model characterizes the behavior of effect size estimates in the absence of selection. The selection model describes the action of selective forces determining which estimates are observed.

Effect Size Model

The present paper extends the Hedges (1992) model by adding linear predictors to the model for unselected effects. One can imagine a situation in which heterogeneity of effects could produce a funnel plot that closely resembles the plot of a group of studies exhibiting publication bias. If, for example, the data set included a number of studies with small sample sizes that, due to some identifiable factor, had legitimately larger effect estimates, that group could mimic the skewed upper tail of a biased funnel plot. Such a situation is plausible; fully randomized experiments, for example, often have smaller sample sizes, and might be expected to yield different results than studies using quasi-experimental designs. The addition of a linear regression model for effects would distinguish such variability from general heterogeneity, as long as the regression model was correctly specified. Moreover, apart from such special cases in which heterogeneity can be confused with selection bias, there may be substantive research questions for which a linear model is appropriate; the sample analysis of the efficacy of psychotherapy data presented below is an example of such a case.

In the absence of selection, one can represent the distribution of sample effects using a linear random-effects model. Let $Y_1, Y_2, \ldots, Y_n$ be variables representing study outcomes (effect size estimates), such that

$$Y_i \sim \mathrm{N}(\delta_i, \sigma_i^2),$$

where σ_i^2 is known, and δ_i is an unknown parameter. Let δ_i be distributed so that

$$\delta_i \sim \mathrm{N}(\Delta_i, \sigma^2),$$

where σ^2 is an unknown variance component, and Δ_i is a function of linear predictors. Specifically, let $\Delta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$, or, in matrix form, $\Delta = \mathbf{X}\boldsymbol{\beta}$, where

$$\mathbf{X}_{n\times p} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}$$

is a matrix of known predictors, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$, is a p-dimensional vector of unknown regression coefficients. Then

$$Y_i \sim \mathrm{N}(\Delta_i, \sigma_i^2 + \sigma^2).$$

The observed statistic Y_i from study i tests the null hypothesis that $\delta_i = 0$ through the test statistic $Z_i = Y_i/\sigma_i$. The one-tailed p-value associated with that test is $p_i = 1 - \Phi(Z_i)$ (assuming that the positive tail is the one of interest), where $\Phi(t)$ denotes the standard normal cumulative distribution function.

Selection Model

The model for selection, following Hedges (1992), describes the probability that an estimate with a particular p-value is observed. The probability is described by a step function over several intervals, with the boundaries of the intervals determined a priori. Boundaries are set at p-values that are important in the mind of the typical researcher. (The literature has shown that psychological researchers tend to consider a result to be much more conclusive if its p-value is just below one of the conventional levels of significance; see, e.g., Rosenthal & Gaito, 1963, 1964; Nelson, Rosenthal, & Rosnow, 1986). Cutpoints at other locations (e.g., $p = .20$, $p = .30$) are added to allow the step function to approximate a putative continuous weight function as closely as possible over the region of probabilities where step discontinuities are not expected. A boundary at .50 is included, to represent the point at which effect estimates become negative. One departure from Hedges' original formulation of the model is that p-values here are one-tailed. Empirical work with these models has suggested that in some domains such as validity generalization studies, funnel plots are more consistent with a one-tailed than a two-tailed pattern of selection (Vevea, Clements & Hedges, 1993). Moreover, as the population effect grows larger, the contribution of the negative tail to the total distribution becomes negligible, so that in many cases, one-tailed and two-tailed selection models yield essentially equivalent results.

The likelihood that an effect from a study with a one-tailed p-value of p_i is observed is represented by a weight function, $\mathrm{w}(p_i)$. Consider a weight function with k intervals of constancy. Denote the left and right endpoints of the j-th such interval by a_{j-1} and a_j, respectively. Let $a_0 = 0$ and $a_k = 1$. If a study has a one-tailed p-value that falls within the j-th such interval, denote its weight by ω_j. If we assume that the weight functions, as functions of p, will be the same for all studies, then

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

The weight function may be defined equivalently as a function of the individual study's effect size, Y_i, and its conditional variance, σ_i^2:

$$\mathrm{w}(Y_i, \sigma_i^2) = \begin{cases} \omega_1 & \text{if } -\sigma_i \Phi^{-1}(a_1) < Y_i \leq \infty; \\ \omega_j & \text{if } -\sigma_i \Phi^{-1}(a_j) < Y_i \leq -\sigma_i \Phi^{-1}(a_{j-1}); \\ \omega_k & \text{if } -\infty < Y_i \leq -\sigma_i \Phi^{-1}(a_{k-1}), \end{cases}$$

where $\Phi^{-1}(p)$ is the inverse normal cumulative distribution function evaluated at p. The number of studies that were present prior to censorship is unknown; hence, the weights are relative, not absolute. This indeterminacy is overcome by constraining the weight for the first p-value interval to be 1.0. A weight for another interval of 0.5, for

example, would therefore indicate that studies with p-values in that interval were only half as likely to be observed as studies in the first interval. Similarly, an estimated weight of 2.0 for another interval would indicate that studies from the new interval were twice as likely to survive the censorship process as were studies from the first interval.

The weighted probability density function of Y_i given the weight function $w(Y_i, \sigma_i^2)$ and the parameters β, σ^2, and $\omega = (\omega_1, \ldots, \omega_k)'$ is

$$f(Y_i|\beta, \sigma^2, \omega) = \frac{w(Y_i, \sigma_i^2)\, \phi\left(\frac{Y_i - \Delta_i}{\eta_i}\right)}{\eta_i A_i(\Delta_i, \eta_i^2, \omega)},$$

where

$$A_i(\Delta_i, \eta_i^2, \omega) = \frac{1}{\eta_i} \int_{-\infty}^{\infty} w(Y_i, \sigma_i^2)\, \phi\left(\frac{Y_i - \Delta_i}{\eta_i}\right) dY_i,$$

$\Delta_i = X_i\beta$, $\phi(z)$ is the standard normal density function evaluated at z, and $\eta_i^2 = \sigma_i^2 + \sigma^2$. The integral A_i may be expressed as the sum of the integrals over the regions where the weight function is constant:

$$A_i(\Delta_i, \eta_i^2, \omega) = \sum_{j=1}^{k} \omega_j B_{ij}(\Delta_i, \sigma^2),$$

where $B_{ij}(\Delta_i, \sigma^2)$ is the probability that a normally distributed random variable with mean Δ_i and variance η_i^2 will be assigned a weight of ω_j. That is,

$$B_{ij} = \begin{cases} 1 - \Phi\left(\frac{b_{i1} - \Delta_i}{\eta_i}\right) & \text{if } j = 1; \\ \Phi\left(\frac{b_{i,j-1} - \Delta_i}{\eta_i}\right) - \Phi\left(\frac{b_{ij} - \Delta_i}{\eta_i}\right) & \text{if } 1 < j < k; \\ \Phi\left(\frac{b_{i,k-1} - \Delta_i}{\eta_i}\right) & \text{if } j = k, \end{cases}$$

where b_{ij} denotes the left endpoint of the interval of Y_i values assigned weight ω_j in the i-th study; that is, $b_{ij} = -\sigma_i \Phi^{-1}(a_j)$.

Assuming that the studies are independent, the joint likelihood for the data $Y = (Y_1, \ldots, Y_n)'$ is the product of the individual likelihoods:

$$\ell(\beta, \sigma^2, \omega) = \prod_{i=1}^{n} \frac{w(Y_i, \sigma_i^2)\, \phi\left(\frac{Y_i - \Delta_i}{\eta_i}\right)}{\eta_i A_i(\Delta_i, \eta_i^2, \omega)},$$

and the log-likelihood is proportional to

$$L = \sum_{i=1}^{n} \log w(Y_i, \sigma_i^2) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{Y_i - \Delta_i}{\eta_i}\right)^2 - \sum_{i=1}^{n} \log \eta_i - \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \omega_j B_{ij}(\Delta_i, \sigma^2)\right).$$

Estimation and Large Sample Standard Errors

Estimates of the model's parameters may be obtained by simultaneously solving their likelihood equations. The likelihood equations for the unconstrained weights $(\omega_2, \ldots, \omega_k)'$ are

$$\frac{\partial L}{\partial \omega_j} = \frac{\mathrm{c}(j)}{\omega_j} - \sum_{i=1}^{n} \frac{B_{ij}}{\sum_{m=1}^{k} \omega_m B_{im}} = 0,$$

where $\mathrm{c}(j)$ is the count of studies assigned the weight ω_j. The equations for the linear predictors are

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij} \left(\frac{Y_i - \Delta_i}{\eta_i^2} \right) - \sum_{i=1}^{n} \frac{\sum_{m=1}^{k} \omega_m \dfrac{\partial B_{im}}{\partial \beta_j}}{\sum_{m=1}^{k} \omega_m B_{im}} = 0,$$

where

$$\frac{\partial B_{im}}{\partial \beta_j} = \begin{cases} \dfrac{X_{ij}}{\eta_i} \phi\left(\dfrac{b_{i1} - \Delta_i}{\eta_i} \right) & \text{if } m = 1; \\ \dfrac{X_{ij}}{\eta_i} \left(\phi\left(\dfrac{b_{im} - \Delta_i}{\eta_i} \right) - \phi\left(\dfrac{b_{i,m-1} - \Delta_i}{\eta_i} \right) \right) & \text{if } 1 < m < k; \\ \dfrac{-X_{ij}}{\eta_i} \phi\left(\dfrac{b_{i,k-1} - \Delta_i}{\eta_i} \right) & \text{if } m = k. \end{cases}$$

The equation for the variance component is

$$\frac{\partial L}{\partial \sigma^2} = \sum_{i=1}^{n} \frac{(Y_i - \Delta_i)^2}{2\eta_i^4} - \sum_{i=1}^{n} \frac{1}{2\eta_i^2} - \sum_{i=1}^{n} \frac{\sum_{m=1}^{k} \omega_m \dfrac{\partial B_{im}}{\partial \sigma^2}}{\sum_{m=1}^{k} \omega_m B_{im}} = 0,$$

where

$$\frac{\partial B_{im}}{\partial \sigma^2} = \begin{cases} \dfrac{b_{i1} - \Delta_i}{2\eta_i^3} \phi\left(\dfrac{b_{i1} - \Delta_i}{\eta_i} \right) & \text{if } m = 1; \\ \dfrac{b_{im} - \Delta_i}{2\eta_i^3} \phi\left(\dfrac{b_{im} - \Delta_i}{\eta_i} \right) - \dfrac{b_{i,m-1} - \Delta_i}{2\eta_i^3} \phi\left(\dfrac{b_{i,m-1} - \Delta_i}{\eta_i} \right) & \text{if } 1 < m < k; \\ \dfrac{b_{i,k-1} - \Delta_i}{2\eta_i^3} \phi\left(\dfrac{b_{i,k-1} - \Delta_i}{\eta_i} \right) & \text{if } m = k. \end{cases}$$

The likelihood equations are solved using the Newton-Raphson algorithm. Successive approximations to the maximum likelihood estimate of the parameter vector $\boldsymbol{\xi} = (\beta_0, \ldots, \beta_p, \sigma^2, \omega_2, \ldots, \omega_k)'$ are given by

$$\xi_{m+1} = \xi_m - \left[\frac{\partial^2 L}{\partial \xi^2}\right]^{-1}_{\xi_m}\left[\frac{\partial L}{\partial \xi}\right]_{\xi_m},$$

where the postmultiplier is the vector obtained by stacking the first derivatives given above, and the premultiplier is the inverse Hessian matrix (i.e., the inverse of the matrix of partial second derivatives and cross derivatives of the log-likelihood with respect to the parameter vector); both are evaluated at the current value of ξ. The algorithm is allowed to iterate until subsequent changes in parameter values are small and first derivatives are near zero.

Asymptotic sampling variances of the estimates are available from the diagonal of the inverse Hessian matrix after estimation has converged. The partial second derivatives that make up the Hessian matrix are given in the appendix.

Application to Correlation Coefficients and Standardized Mean Differences

The likelihood equations and other derivatives require that the conditional sampling variances (σ_i^2) of the estimates from individual studies be known. The form these variances take depends upon what type of effects the studies estimate. When the effects are correlation coefficients, as is the case in validity generalization studies, it may be desirable to transform them by Fisher's Z-transformation so that the individual variances are then $1/(n_i - 3)$, where n_i is the sample size of the i-th study. Note that these variances do not depend on the correlation parameter.

When the effects represent standardized differences between means, and the ratio of the control group's sample size to the experimental group's sample size is constant across studies, a variance stabilizing transformation of the effect is known (see Hedges & Olkin, 1985). However, the need for equal ratios of sample sizes is unlikely to be met in any data set large enough to support adequate estimation of the weight function employed in the present model, and the behavior of the transformation when sample size ratios are unequal has not been studied. An alternative is to approximate the conditional sampling variances by

$$s_i^2 = \frac{2}{n_i}\left(1 + \frac{d_i^2}{4}\right),$$

where n_i is the square mean root of the two sample sizes. Although these variances do depend on the estimated effect magnitude, when that effect is small (as it typically is in social science research), its contribution to the sampling variance is negligible; thus the variances produced by this formula may be treated as essentially known.

Likelihood Ratio Tests of Nested Models

In addition to providing adjusted estimates of the linear predictors and the variance component, the model allows the construction of likelihood ratio tests for differences in fit among different specifications with different constrained parameters.

Testing the Effect Size Model

One application of such tests is the comparison of different linear effect size models. For example, one might wish to compare a model that assumes a common population effect across all studies ($\Delta_i = \beta_0$) with a model that adds an effect for some study characteristic ($\Delta_i = \beta_0 + \beta_1 X_{i1}$). One may view the first model as a special case of the second with β_1 constrained to be zero. Then minus two times the difference be-

tween the maximum of log-likelihoods under the respective models has approximately a chi-square distribution with degrees of freedom equal to the number of constrained parameters in the simpler model—here, one (see, e.g., Kendall & Stuart, 1979). In this example, the test is asymptotically equivalent to a test of the null hypothesis that $\beta_1 = 0$ based on the Z-statistic obtained by dividing the estimate of β_1 by its standard error. However, one could compare the first model ($\Delta_i = \beta_0$) with a more complex model (e.g., $\Delta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$); now the likelihood ratio chi-square statistic has three degrees of freedom, and constitutes a test of the significance of the entire linear model.

Testing the Selection Model

Another useful application of such tests is to compare the model that estimates a set of weights representing the selection function to the model that constrains those weights to be one. Since the first of the k weights is always constrained to be one, the likelihood ratio statistic will have $k - 1$ degrees of freedom, and will test the improvement in fit obtained by adding the selection model to the linear random-effects model. When the pattern of estimated weights appears to reflect a reasonable process of differential selection, the statistic may be taken to be a test for the presence of publication bias. In principle, the test will detect any deviation from uniformity in the weights. The interpretation of that non-uniformity must be made in light of the pattern of the weights. For example, the classic "bias toward significant results" pattern would have larger weights for smaller p-values. Another plausible pattern might be "bias toward positive results", which would be exhibited if weights were relatively constant for p-values below .5 and smaller for p-values above .5. Other patterns of weights that could be interpreted substantively are also possible.

The weight function we have proposed depends only on the one-tailed p-value associated with the effect size estimate, and is the same for all studies. It is possible to conceive that the selection model (and hence the weight function) also depends on qualitative characteristics of the studies. For example, one might imagine that the selection model is different for randomized experiments than for quasi-experiments (i.e., because they are presumed to be more valid, randomized experiments are more likely to be published than quasi-experiments yielding the same p-value). Such a case might be handled by estimating a different weight function for each of the two groups of studies. More generally, one might posit that the weights depend on a linear model of study characteristics in addition to p-values. Such a selection model might be estimated with a suitable modification of the techniques presented in this paper.

An Application of the Linear Model for Effects

One of the best-known series of meta-analytic studies is the analysis performed by Smith, Glass and Miller (1980) on the efficacy of psychotherapy. Although techniques for meta-analysis have improved since the publication of their results, and the data set they analyzed may have been flawed (e.g., there were effects that would probably be considered impossibly large by most analysts), the study remains an important one, and the conclusions continue to be of interest. A large number of the studies that the psychotherapy analysis included consisted of either behavioral therapies or systematic desensitization treatments of phobias. Phobics were subdivided into "true" or "complex" phobics, who exhibited multiple phobias, and "simple" or monosymptomatic phobics. The analysis presented here is limited to those studies, and considers only effect estimates that are equivalent to standardized mean differences. Because of the outlying effects in the original data set (one study with an effect of 25.33 was present),

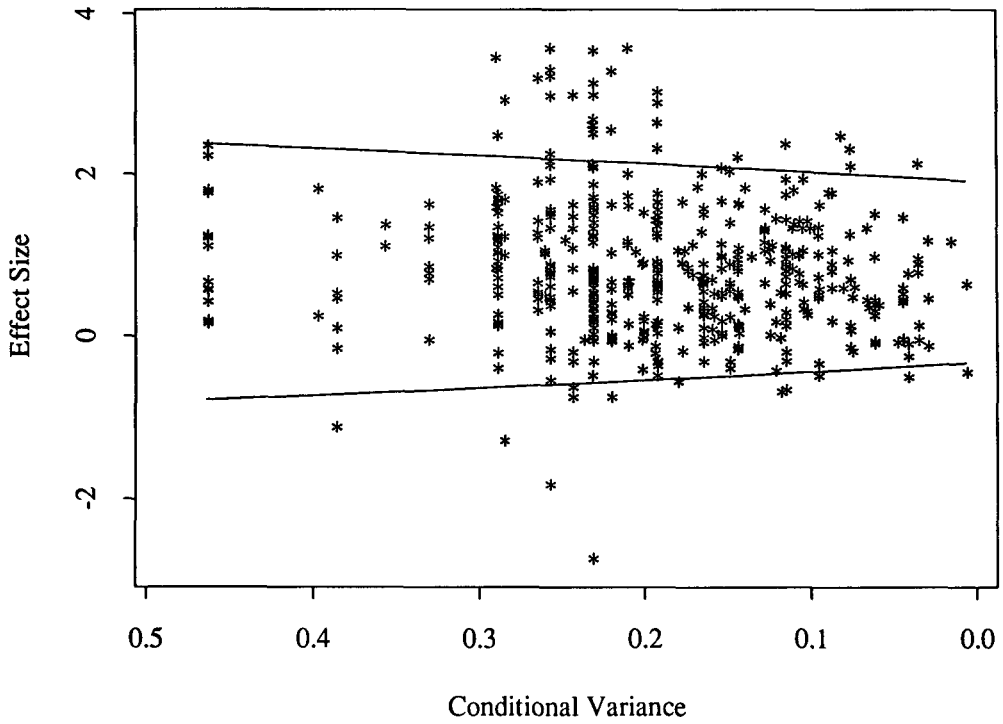

FIGURE 3.
Modified funnel plot of psychotherapy effects.

studies with effects higher than 4.0 were deleted; only five data points were lost in that step. After those criteria had been applied, 489 studies remained, of which 216 employed behavioral treatments, and 273 were studies of desensitization therapies. Of the 216 behavioral studies, 130 involved true phobics, and 86 involved simple phobics. The desensitization studies included 59 with true phobics, and 214 with simple phobics. Figure 3 presents the modified funnel plot of these studies; the plot clearly shows the typical asymmetry associated with one-tailed selection. Since the ratio of the control group's sample size to the treatment group's sample size is not constant across the studies in the data set, and the estimates of effect size tend not to be large, we approximate the conditional sampling variances without employing the stabilizing transformation mentioned above. The conventional random-effects model estimates the variance component to be 0.456, and the common effect to be 0.800.

Two linear models were considered. The more complex model proposed is a full factorial analysis of variance (ANOVA) with interaction:

$$\Delta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i},$$

where

$$X_{1i} = \begin{cases} 0 & \text{if treatment is behavioral;} \\ 1 & \text{if treatment is desensitization,} \end{cases}$$

and

$$X_{2i} = \begin{cases} 0 & \text{if phobia is complex;} \\ 1 & \text{if phobia is simple.} \end{cases}$$

TABLE 1

Psychotherapy Outcomes, Not Adjusted for Publication Bias

Therapy Type	Condition Treated	Predicted Mean	Standard Error of Prediction
Behavioral	Complex Phobia	0.628	0.058
	Simple Phobia	0.900	0.079
Desensitization	Complex Phobia	0.704	0.092
	Simple Phobia	0.857	0.049

The factorial ANOVA model results in a variance component estimate of 0.280 (s.e. = 0.032) when selection is not included in the model, and 0.375 (s.e. = 0.069) when selection is accounted for. Note that these estimates are considerably lower than the single-common-effect model's estimate; one may think of the additional linear model as accounting for some of variability that is incorporated into the original variance component. (When selection is not accounted for, the reduction associated with the added linear model is 0.456 − 0.280 = 0.176, or 39 percent of the total variance component.) Estimates of the predictors in the factorial model without accounting for selection are $\beta_0 = 0.628$ (s.e. = 0.058); $\beta_1 = 0.076$ (s.e. = 0.108); $\beta_2 = 0.272$ (s.e. = 0.097); and $\beta_3 = -0.119$ (s.e. = 0.142).

Estimated means for the four possible conditions are given in Table 1. Standard errors for the means were constructed from the parameter covariance matrix, and are included in the table. The likelihood ratio test statistic for the added linear model was 28.94 on 3 degrees of freedom, $p < .0001$, indicating that the factorial ANOVA model clearly fits better than the single-common-effect model. When the selection component of the model was added, p-value intervals were set to .000–.001, .001–.005, .005–.010, .010–.020, .020–.050, .050–.100, .100–.200, .200–.300, .300–.500, and .500–1.000. The weight for the first interval was, as usual, constrained to 1.0; estimated weights for the remaining intervals were 1.487 (s.e. = 0.308), 1.253 (s.e. = 0.332), 1.685 (s.e. = 0.425), 1.169 (s.e. = 0.313), 1.152 (s.e. = 0.340), 0.894 (s.e. = 0.291), 0.834 (s.e. = 0.310), 0.739 (s.e. = 0.291), and 0.572 (s.e. = 0.292). The weights for nonsignificant (i.e. $p > .05$) intervals decrease monotonically as the p-value increases. The weights that were greater than 1.0 indicate that studies in those intervals were more likely to be included than were studies in the first (most significant) interval; that can occur only if there are fewer studies than expected in the first interval. Thus, while a reasonable p-value-related selection process appears to have occurred, it seems that studies with extremely large effect estimates may also have been censored to a greater degree than could be accounted for by our culling of effects greater than 4.0. The likelihood ratio test for the addition of the weights to the model was 22.066 on 9 degrees of freedom, $p = .0087$,

TABLE 2

Psychotherapy Outcomes, Adjusted for Publication Bias

Therapy Type	Condition Treated	Predicted Mean	Standard Error of Prediction
Behavioral	Complex Phobia	0.482	0.133
	Simple Phobia	0.767	0.151
Desensitization	Complex Phobia	0.531	0.173
	Simple Phobia	0.727	0.147

indicating that the weight function substantially improved the fit of the model. Adjusted estimates of the linear model's predictors are $\beta_0 = 0.482$ (s.e. = 0.140); $\beta_1 = 0.049$ (s.e. = 0.130); $\beta_2 = 0.285$ (s.e. = 0.119); and $\beta_3 = -0.089$ (s.e. = 0.172). Adjusted means for the four possible treatment/phobia conditions are given in Table 2. Those adjustments represent reductions of 15 percent for either treatment of simple phobias, and 23 or 25 percent for behavioral or desensitization treatment of complex phobias. Note, however, that the standard errors of the predicted means are two to three times as large as they were before the weight function was estimated.

The standard errors of the coefficients of the factorial ANOVA model suggest that little would be lost by dropping the main effect of treatment type and the interaction, leaving the model $\Delta_i = \beta_0 + \beta_1 X_2$, where X_2 is defined as before—zero if the phobia is complex, or one if it is simple. Likelihood ratio tests both with and without the selection model confirm that suggestion. Without the selection model, the test statistic for the effect of dropping the parameters is 0.710 on two degrees of freedom, $p = .7012$; with the selection model, the statistic is 0.439 on two degrees of freedom, $p = .8029$. The coefficient estimates for this simpler model are $\beta_0 = 0.650$ (s.e. = 0.049), $\beta_1 = 0.219$ (s.e. = 0.064) before the selection model is applied, and $\beta_0 = 0.496$ (s.e. = 0.136), $\beta_1 = 0.242$ (s.e. = 0.078) when the weight function is added. (The estimated weights are virtually identical to those of the factorial model.) The effect estimate for either treatment of simple phobias, then, moves from 0.869 to 0.738, a 15 percent reduction; the estimate for complex phobias is reduced by 25 percent, from 0.650 to 0.496.

One note of caution is in order. It would be wrong to conclude that the lack of need for an interaction term or therapy-type main effect in the model implies that behavioral therapies and desensitization therapies are equally effective against either simple or complex phobias. Recall that the proportion of each type of therapy that was applied to simple or complex phobias differed. Of the studies employing behavioral therapies, 60 percent treated complex phobias, whereas 78 percent of the desensitization studies

treated simple phobias. Thus the situation is similar to an unbalanced ANOVA: the two main effects are confounded.

Conclusions

The model described and illustrated above shows one way to approach the problem of publication bias. Any model requires that assumptions be made about the nature of the problem. The assumptions of this particular model may be unusually strong, in that they include definite statements about the form of the distribution of the random effects (normality) and somewhat weaker statements about the form of the weight function. Those statements, however, appear to be plausible, and the model has performed well in simulations, even under rather extreme violations of distributional assumptions (Hedges & Vevea, 1993). In particular, the model's ability to reduce the bias of effect estimates when censorship has occurred appears to be quite robust to violations in the form of the distribution of random effects so long as the between-studies variance component σ^2 is not large compared with the conditional variances $\sigma_1^2, \ldots, \sigma_n^2$. Indeed, the procedure described here should be more robust than those studied by Hedges and Vevea (1993) because it models some of the between-study variation as due to study characteristics, and so reduces the magnitude of between-study variation. Nevertheless, the results should be treated with caution. Note, for example, that 95 percent confidence intervals for the predicted means after estimation of the weight function (Table 2) would in all cases include the predicted means estimated without the selection model (Table 1). It might be more appropriate, then, to interpret the psychotherapy results as indicating that bias is probably present and may be substantial, so that the conventional effect estimates should be regarded with skepticism, rather than placing great confidence in the exact values of model-adjusted effects.

Appendix

The requisite derivatives for the diagonal of the Hessian matrix are

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

$$\frac{\partial^2 L}{\partial \beta_j^2} = \left(\frac{\sum_{m=1}^{k} \omega_m \frac{\partial B_{im}}{\partial \beta_j}}{\sum_{m=1}^{k} \omega_m B_{im}} \right)^2 - \sum_{i=1}^{n} \frac{\sum_{m=1}^{k} \omega_m \frac{\partial^2 B_{im}}{\partial \beta_j^2}}{\sum_{m=1}^{k} \omega_m B_{im}} - \sum_{i=1}^{n} \frac{X_{ij}^2}{\eta_i^2},$$

and

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

$$+ \sum_{i=1}^{n} \left(\left(\frac{\sum_{m=1}^{k} \omega_m \frac{\partial B_{im}}{\partial (\sigma^2)^2}}{\sum_{m=1}^{k} \omega_m B_{im}} \right)^2 \right),$$

where the first partial derivatives of B_{im} are the same as before, and the second partial derivatives are

$$\frac{\partial^2 B_{im}}{\partial \beta_j^2} = \begin{cases} X_{ij}^2 \dfrac{b_{i1} - \Delta_i}{\eta_i^3}\, \phi\left(\dfrac{b_{i1} - \Delta_i}{\eta_i}\right) & \text{if } m = 1; \\ X_{ij}^2 \left(\dfrac{b_{im} - \Delta_i}{\eta_i^3}\, \phi\left(\dfrac{b_{im} - \Delta_i}{\eta_i}\right) - \dfrac{b_{i,m-1} - \Delta_i}{\eta_i^3}\, \phi\left(\dfrac{b_{i,m-1} - \Delta_i}{\eta_i}\right)\right) & \text{if } 1 < m < k; \\ -X_{ij}^2 \dfrac{b_{i,k-1} - \Delta_i}{\eta_i^3}\, \phi\left(\dfrac{b_{i,k-1} - \Delta_i}{\eta_i}\right) & \text{if } m = k, \end{cases}$$

and

$$\frac{\partial^2 B_{im}}{\partial (\sigma^2)^2} = \begin{cases} \dfrac{(b_{i1} - \Delta_i)^3}{4\eta_i^7} - \dfrac{3(b_{i1} - \Delta_i)}{4\eta_i^5}\, \phi\left(\dfrac{b_{i1} - \Delta_i}{\eta_i}\right) & \text{if } m = 1; \\ \left(\dfrac{(b_{im} - \Delta_i)^3}{4\eta_i^7} - \dfrac{3(b_{im} - \Delta_i)}{4\eta_i^5}\right) \phi\left(\dfrac{b_{i,m-1} - \Delta_i}{\eta_i}\right) - \\ \left(\dfrac{(b_{i,m-1} - \Delta_i)^3}{4\eta_i^7} - \dfrac{3(b_{i,m-1} - \Delta_i)}{4\eta_i^5}\right) \phi\left(\dfrac{b_{i,m-1} - \Delta_i}{\eta_i}\right) & \text{if } 1 < m < k; \\ -\left(\dfrac{(b_{i,k-1} - \Delta_i)^3}{4\eta_i^7} - \dfrac{3(b_{i,k-1} - \Delta_i)}{4\eta_i^5}\right) \phi\left(\dfrac{b_{i,k-1} - \Delta_i}{\eta_i}\right) & \text{if } m = k. \end{cases}$$

The off-diagonal elements of the matrix are

$$\frac{\partial^2 L}{\partial \omega_j \partial \omega_\ell} = \sum_{i=1}^{n} \frac{B_{ij} B_{i\ell}}{\left(\sum_{m=1}^{k} \omega_m B_{im}\right)^2},$$

$$\frac{\partial^2 L}{\partial \omega_j \partial \beta_\ell} = \sum_{i=1}^{n} \left(\frac{B_{ij} \left(\sum_{m=1}^{k} \omega_m \dfrac{\partial B_{im}}{\partial \beta_\ell} \right)}{\left(\sum_{m=1}^{k} \omega_m B_{im}\right)^2} - \frac{\dfrac{\partial B_{ij}}{\partial \beta_\ell}}{\sum_{m=1}^{k} \omega_m B_{im}} \right),$$

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

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_\ell} = \sum_{i=1}^{n} \frac{\left(\sum_{m=1}^{k} \omega_m \dfrac{\partial B_{im}}{\partial \beta_j} \right) \left(\sum_{m=1}^{k} \omega_m \dfrac{\partial B_{im}}{\partial \beta_\ell} \right)}{\left(\sum_{m=1}^{k} \omega_m B_{im}\right)^2}$$

$$- \sum_{i=1}^{n} \frac{\sum_{m=1}^{k} \omega_m \dfrac{\partial^2 B_{im}}{\partial \beta_j \partial \beta_\ell}}{\sum_{m=1}^{k} \omega_m B_{im}} - \sum_{i=1}^{n} \frac{X_{ij} X_{i\ell}}{\eta_i^2},$$

and

$$\frac{\partial^2 L}{\partial\beta_j\partial\sigma^2} = \sum_{i=1}^{n} \frac{\left(\sum_{m=1}^{k} \omega_m \frac{\partial B_{im}}{\partial\beta_j}\right)\left(\sum_{m=1}^{k} \omega_m \frac{\partial B_{im}}{\partial\sigma^2}\right)}{\left(\sum_{m=1}^{k} \omega_m B_{im}\right)^2}$$

$$- \sum_{i=1}^{n} \frac{\sum_{m=1}^{k} \omega_m \frac{\partial^2 B_{im}}{\partial\beta_j\partial\sigma^2}}{\sum_{m=1}^{k} \omega_m B_{im}} - \sum_{i=1}^{n} X_{ij} \frac{Y_i - \Delta_i}{\eta_i^4},$$

where

$$\frac{\partial^2 B_{im}}{\partial\beta_j\partial\beta_\ell} = \begin{cases} X_{ij}X_{i\ell} \dfrac{b_{i1} - \Delta_i}{\eta_i^3} \phi\left(\dfrac{b_{i1} - \Delta_i}{\eta_i}\right) & \text{if } m = 1; \\ X_{ij}X_{i\ell}\left[\dfrac{b_{im} - \Delta_i}{\eta_i^3} \phi\left(\dfrac{b_{im} - \Delta_i}{\eta_i}\right) - \dfrac{b_{i,m-1} - \Delta_i}{\eta_i^3} \phi\left(\dfrac{b_{i,m-1} - \Delta_i}{\eta_i}\right)\right] & \text{if } 1 < m < k; \\ -X_{ij}X_{i\ell} \dfrac{b_{i,k-1} - \Delta_i}{\eta_i^3} \phi\left(\dfrac{b_{i,k-1} - \Delta_i}{\eta_i}\right) & \text{if } m = k, \end{cases}$$

and

$$\frac{\partial^2 B_{im}}{\partial\beta_j\partial\sigma^2} = \begin{cases} \dfrac{X_{ij}}{2\eta_i^3} \phi\left(\dfrac{b_{i1} - \Delta_i}{\eta_i}\right)\left[\left(\dfrac{b_{i1} - \Delta_i}{\eta_i}\right)^2 - 1\right] & \text{if } m = 1; \\ \dfrac{X_{ij}}{2\eta_i^3} \phi\left(\dfrac{b_{im} - \Delta_i}{\eta_i}\right)\left[\left(\dfrac{b_{im} - \Delta_i}{\eta_i}\right)^2 - 1\right] - & \\ \dfrac{X_{ij}}{2\eta_i^3} \phi\left(\dfrac{b_{i,m-1} - \Delta_i}{\eta_i}\right)\left[\left(\dfrac{b_{i,m-1} - \Delta_i}{\eta_i}\right)^2 - 1\right] & \text{if } 1 < m < k; \\ \dfrac{-X_{ij}}{2\eta_i^3} \phi\left(\dfrac{b_{i,k-1} - \Delta_i}{\eta_i}\right)\left[\left(\dfrac{b_{i,k-1} - \Delta_i}{\eta_i}\right)^2 - 1\right] & \text{if } m = k. \end{cases}$$

References

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges, *The handbook of research synthesis* (pp. 399–409). New York: Russell Sage Foundation.

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data (with discussion). *Journal of the Royal Statistical Society, Series A, 151*, 419–463.

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

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Coursol, A., & 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., & Williams, M. (1984). Discussion on meta-analysis and selective publication bias. *American Psychologist, 39*, 75–78.

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

Dickersin, K., Min, Y-I, & Meinert, C. L. (1991). The fate of controlled trials funded by the NIH in 1979. *Controlled Clinical Trials, 12,* 634.

Dickersin, K., Min, Y-I, & Meinert, C. L. (1992). Factors influencing the publication of research results: Followup of applications submitted to two institutional review boards. *Journal of the American Medical Association, 267,* 374–378.

Easterbrook, P. J., Berlin, J. A., Gopalan, R., & Matthews, D. R. (1991). Publication bias in clinical research. *Lancet, 337,* 867–872.

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

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. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science, 7,* 246–255.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hedges, L. V., & Vevea, J. L. (1993). *Estimating effect size under publication bias: Small sample properties and robustness of a selection model.* Manuscript submitted for publication.

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

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics. Volume 2, Inference and relationship* (4th ed.). London and High Wycombe: Charles Griffin and Company.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31,* 107–112.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

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

National Research Council (1992). *Combining information: Statistical issues and research opportunities.* Washington, DC: National Academy Press.

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

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

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports, 4,* 570.

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

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: The Johns Hopkins University Press.

Sterling, T. C. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Vevea, J. L., Clements, N. C., & Hedges, L. V. (1993). Assessing the effects of selection bias on validity data for the general aptitude test battery. *Journal of Applied Psychology, 78,* 981–987.

White, K. R. (1982). The relation between socioeconomic status and achievement. *Psychological Bulletin, 31,* 461–481.

Manuscript received 5/13/94
Final version received 8/29/94