Psychological Methods
2005, Vol. 10, No. 4, 428–443

1082-989X/05/$12.00 DOI: 10.1037/1082-989X.10.4.428

Publication Bias in Research Synthesis: Sensitivity Analysis Using A Priori Weight Functions

Jack L. Vevea
University of California, Santa Cruz

Carol M. Woods
Washington University in St. Louis

Publication bias, sometimes known as the "file-drawer problem" or "funnel-plot asymmetry," is common in empirical research. The authors review the implications of publication bias for quantitative research synthesis (meta-analysis) and describe existing techniques for detecting and correcting it. A new approach is proposed that is suitable for application to meta-analytic data sets that are too small for the application of existing methods. The model estimates parameters relevant to fixed-effects, mixed-effects or random-effects meta-analysis contingent on a hypothetical pattern of bias that is fixed independently of the data. The authors illustrate this approach for sensitivity analysis using 3 data sets adapted from a commonly cited reference work on research synthesis (H. M. Cooper & L. V. Hedges, 1994).

Publication bias, or the "file-drawer problem" (Rosenthal, 1979), is the tendency for the availability of research to depend on the results. In a simple (and extreme) case, publication bias might manifest itself if only studies with results that are statistically significant at a conventional level (e.g., $p < .05$ or .01) are published, and all other studies are not published (Rosenthal, 1979). This censorship would result in serious inflation of Type I error rates in the published literature (Denton, 1987, 1990) and possibly in the administration of ineffective or dangerous forms of medical or psychological care (Chalmers, 1990). The suppression of significant effects by researchers who expect null results is another possible (though rarely observed) form of publication bias. Any censoring scheme whereby the outcome of a study influences its availability constitutes publication bias.

There is ample evidence that publication bias exists. Empirical research using various methodologies, and conducted in diverse substantive areas, suggests that statistically significant results are more likely than nonsignificant results to be published and presented (Bozarth & Roberts, 1972; Greenwald, 1975; Smart, 1964; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). Average effect sizes tend to be larger for published versus unpublished research (Glass, Smith, & Barton, 1979, as cited in Smith, 1980; Lipsey & Wilson, 1993). Publication bias also has been evident when researchers track studies reported in summary form (Chalmers et al., 1990; Koren, Graham, Shear, & Einarson, 1989; Weber, Callaham, Wears, Barton, & Young, 1998) or approved by an ethics committee (Cooper, DeNeve, & Charlton, 1997; Dickersin, Min, & Meinert, 1992; Easterbrook, Berlin, Gopalan, & Matthews, 1991; Stern & Simes, 1997) or when authors are surveyed about unpublished work (e.g., Chan, Sacks, & Chalmers, 1982; Coursol & Wagner, 1986; Dickersin, Chan, Chalmers, Sacks, & Smith, 1987; Easterbrook et al., 1991; Misakian & Bero, 1998; Rotton, Foos, Vanmeek, & Levitt, 1995; Scherer, Dickersin, & Langenberg, 1994; Shadish, Doherty, & Montgomery, 1989; Sommer, 1987; Weber et al., 1998). Further, evidence suggests that even when published, nonsignificant results are less accessible, because of less media attention (Koren & Klein, 1991), lower likelihood of publication in English (Egger, Zellwager-Zähner, et al., 1997; Grégoire, Derderian, & Le Lorier, 1995), less chance of publication in widely read journals (Easterbrook et al., 1991; Simes, 1987), and greater lag time to publication (Stern & Simes, 1997).

In this article, we review existing statistical and graphical methods for addressing the problem of publication bias in meta-analysis. One problem that emerges in this survey is that the most versatile methods also tend to involve complicated models that are difficult to estimate with typical meta-analytic data sets. We propose a modification of one such model that relieves this problem by adopting a sensitivity-analysis approach, and we illustrate the method using three familiar archival data sets.

Jack L. Vevea, Department of Psychology, University of California, Santa Cruz; Carol M. Woods, Department of Psychology, Washington University in St. Louis.

Additional materials are on the Web at http://dx.doi.org/10.1037/1082=989X.10.4.428.supp

This work was partially supported by National Institute on Aging Grant AG024771,"Software for Meta-Regression, under the SBI program.

Correspondence concerning this article should be addressed to Jack L. Vevea, Department of Psychology, University of California, Room 261 Social Sciences II, 1156 High Street, Santa Cruz, CA 95064. E-mail: jvevea@ucsc.edu

Publication Bias and Meta-Analysis

Although publication bias is thought to affect published research in general, and thus has consequences for the casual consumer of research results (Begg, 1985; Dawid & Dickey, 1977; Denton, 1987, 1990; Hedges, 1984; Hedges & Olkin, 1985), it is often discussed in the context of quantitative research synthesis or meta-analysis (e.g., Hedges, 1984). It is a particular concern for the meta-analyst, because meta-analytic results depend on the assumption that available studies are a random sample of all those that exist on the topic. If the studies available for synthesis are not representative, the validity of the conclusions is threatened. Given the potentially serious implications of publication bias, a number of authors have suggested strategies for eliminating or preventing bias, as well as statistical methods for detecting and correcting it in the context of meta-analysis.

Eliminating or Preventing Publication Bias

A number of promising ideas have been presented for the elimination and prevention of publication bias in the long term. For example, various concerned researchers have proposed the elimination of hypothesis testing (Begg & Berlin, 1989; Hubbard & Armstrong, 1997; Nunnally, 1960), a priori peer review and failure to publish studies with inadequate sample sizes (Newcombe, 1987), development of a more positive attitude toward nonsignificant results (Greenwald, 1975; Rennie & Flanagin, 1992), and improvement of the peer-review and publication processes (Higginson, 1987; Iyengar & Greenhouse, 1988; Mahoney, 1977; Newcombe, 1987; Smart, 1964; Sterling et al., 1995; Wolff, 1973). However, not all of these ideas are popular, and even if steps were taken, it would be many years before all of the relevant literature one might want to synthesize could be free from the threat of publication bias. Techniques for conducting meta-analyses that involve excluding studies based on sample size (e.g., Begg & Berlin, 1988; Kraemer, Gardner, Brooks, & Yesavage, 1998), registration status (Simes, 1986, 1987), publication status (e.g., Sohn, 1996), or methodological quality (e.g., Greenwald & Russell, 1991) are controversial and may introduce other kinds of bias. Thus, additional ideas for coping with publication bias are needed.

Assessing Publication Bias

Numerous statistical procedures have been developed to test whether a sample of identified studies is biased or to assess the impact of any such bias. Most are based on the assumption that, for a given substantive area, studies with small samples should yield a relatively wide range of effect sizes, whereas studies with large samples should yield an effect near to the population effect size. Thus, if a collection of published effect sizes includes few small studies with small effects, it may be the result of a bias against statistically nonsignificant findings. Bias against significant findings is rarely mentioned in this literature but also could be tested with most of the methods reviewed here.

Fail-safe n and related procedures. One common approach to assessing whether publication bias is a problem in a particular data set is to estimate the number of unidentified studies that would be required in order to alter the conclusions (Rosenthal, 1978). This was the purpose of Rosenthal's (1978, 1979) fail-safe *n*, designed for use with synthesized probability values. The test was based on the assumption that unidentified studies had a one-tailed probability value of .5 and an effect size equal to 0. These two assumptions are problematic because meta-analysts often are interested in effect sizes, rather than probability values, and the missing effect sizes are unlikely to be zero (just as the missing probability values are unlikely to be .5). Accordingly, the fail-safe *n* has been modified for use with effect sizes, which can be nonzero (Ashworth, Osburn, Callender, & Boyle, 1992; Orwin, 1983). Iyengar and Greenhouse (1988) addressed the issue by altering the original formula to reflect the assumption that unpublished studies are a sample from an appropriately truncated normal distribution.

Two newer procedures have been developed, which are similar to the fail-safe *n* but differ because what is estimated is the number of missing studies, rather than the number of studies that would reverse meta-analytic conclusions (Gleser & Olkin, 1996; Silliman, 1997). One technique provides a statistical test for whether the *n* missing studies would overturn conclusions, which is more objective than the fail-safe *n*; however, it also assumes that the null hypothesis is true, and it has not been modified for use with effect sizes rather than probability values (Gleser & Olkin, 1996).

Several weaknesses limit the utility of the procedures that are related to the fail-safe-*n* approach. Some of the approaches simply ask the wrong question for the purposes of most meta-analyses: Combining probability values is less generally relevant than it once was thought to be. Improved methods that focus on effect size may require that researchers guess the effect sizes of file-drawer studies based on published research (or intuition), and there is no way to verify the accuracy of the estimate. The number of studies that are unobserved or that would reverse conclusions is of less interest than the bias present in the data one has (Givens, Smith, & Tweedie, 1997b). The methods cannot incorporate models for heterogeneity and, hence, are likely to mislead when heterogeneity is present. The fail-safe-*n* procedures have been criticized for being atheoretic and frequently subject to misinterpretation. For all of those reasons, as well as others (Becker, in press), the use of methods related to the fail-safe *n* should be avoided.

Funnel plots and related procedures. Another approach to detecting publication bias is to determine, among a group of identified studies, the proportion with both a small sample and a small effect size. One well-known graphical technique is Light and Pillemer's (1984) funnel plot, a scatter plot of effect size graphed against sample size (or an expression of sampling uncertainty such as standard error or conditional variance, e.g., Vevea & Hedges, 1995) that is centered on the true population effect size. If there is no publication bias, the point cloud is funnel shaped, reflecting the greater variability of the effect-size estimates from small studies. In the presence of publication bias, the plot will lack symmetry. If the true population effect size is nonzero, the part of the graph where the effects and the samples are both small is sparse. If the true population effect size is near zero, and publication bias has favored both positive and negative significant results, a plot indicative of publication bias is hollow around effect size zero for all but very large samples. (For large samples, even very small effects may be significant.) If publication bias is one-tailed,[1] and the true effect is zero (or near zero), the plot appears sparse for small-sample studies with effects above zero, and it is also truncated below zero.

Interpretation of funnel plots can be difficult, but some researchers have developed methods aimed at aiding interpretation (Berlin, Begg, & Louis, 1989; Egger, Smith, Schneider, & Minder, 1997), and others have introduced procedures based on the same principles as the funnel plot, but with more objective criteria for the presence of bias (e.g., Begg, 1994; Begg & Mazumdar, 1994; Copas, 1999; Copas & Li, 1997; Wang & Bushman, 1998). However, all of these funnel-plot-related procedures can be misleading when the magnitudes of effects depend on study characteristics. Figure 1 illustrates this phenomenon. The first panel shows an evidently biased simulated meta-analytic data set. The asymmetry may not be immediately apparent: Funnel plots can be subtle. However, a close examination will reveal a tendency for studies with weights in the vicinity of 100 to be large; moreover, the upper tail of the plot stretches further in the positive direction than the lower tail does in the negative direction. The second panel reveals that the apparent bias is an artifact of the superimposition of effects drawn from two distinct populations.[2] Funnel plots, then, are sometimes useful tools for bias detection, but their interpretation is often problematic, and they leave open the question of how to proceed if publication bias is suspected.

Correcting for Publication Bias

Two primary methods are currently available that yield an average effect-size estimate corrected for publication bias. As we explain, the merits of each method depend on the circumstances for their use.

Trim and fill. Following the identification of a biased-looking funnel plot, Duval and Tweedie's (2000a, 2000b) trim-and-fill procedure can be used to estimate an average effect size that is, subject to the assumptions of the method, corrected for publication bias (Duval & Tweedie, 2000a, 2000b; Sutton, Duval, Tweedie, Abrams, & Jones, 2000). Trim and fill is a procedure by which the number (k) of studies missing from the truncated part of a biased funnel plot is estimated. Next, the k studies with the largest effects are reflected onto the negative side of the funnel plot, so that k artificial studies with small effect sizes are added to the data set. A new mean effect size is then calculated, which leads to a new estimate of the number of missing effects. The process iterates until no further changes are observed. Finally, a new effect-size estimate is calculated based on the original data set with the new phantom studies added in. This final estimate is interpreted as the average (weighted) effect-size estimate corrected for publication bias.

Trim and fill is advantageous because it can be applied to small data sets; however, it requires the strict assumption that all suppressed studies are those with the most negative effect sizes. If this assumption is inaccurate, the "corrected" effect-size estimate will be inaccurate. Also, the trim-and-fill method cannot estimate predictive models for effect size (analysis-of-variance- and regression-like models in which effect sizes are the dependent variable), and the method can lead to spurious adjustments for publication bias if effects are heterogeneous (Terrin, Schmid, Lau, & Olkin, 2003).

Weight-function models of publication bias. A second method for correcting bias avoids the assumption of strict determinacy in the order in which effects are missing, but the cost of flexibility is that a large sample (e.g., 100 or more studies)[3] is needed for reliable estimates. In weight-function models of publication bias, the weights represent the process by which some studies are more likely to be published than others, based on study characteristics like statistical significance. Specific assumptions associated with

[1] By *one-tailed* as opposed to *two-tailed*, we do not imply that primary researchers used one-tailed significance tests, but, rather, that rejections of a null hypothesis in a certain direction are more available than both nonsignificant results and rejections of the null hypothesis in the unanticipated direction.

[2] In this article, funnel plots portray effect sizes on the y-axis and fixed-effects weights on the horizontal axis. This is essentially the same as plotting effect size against sample size.

[3] It is often possible to estimate the Vevea and Hedges (1995) weight model with fewer than 100 effect sizes. However, to do so, one must reduce the number of weights used to describe the selection function. Under such circumstances, the resulting statistical estimates of effect size may vary depending on the exact locations of probability value cutpoints.

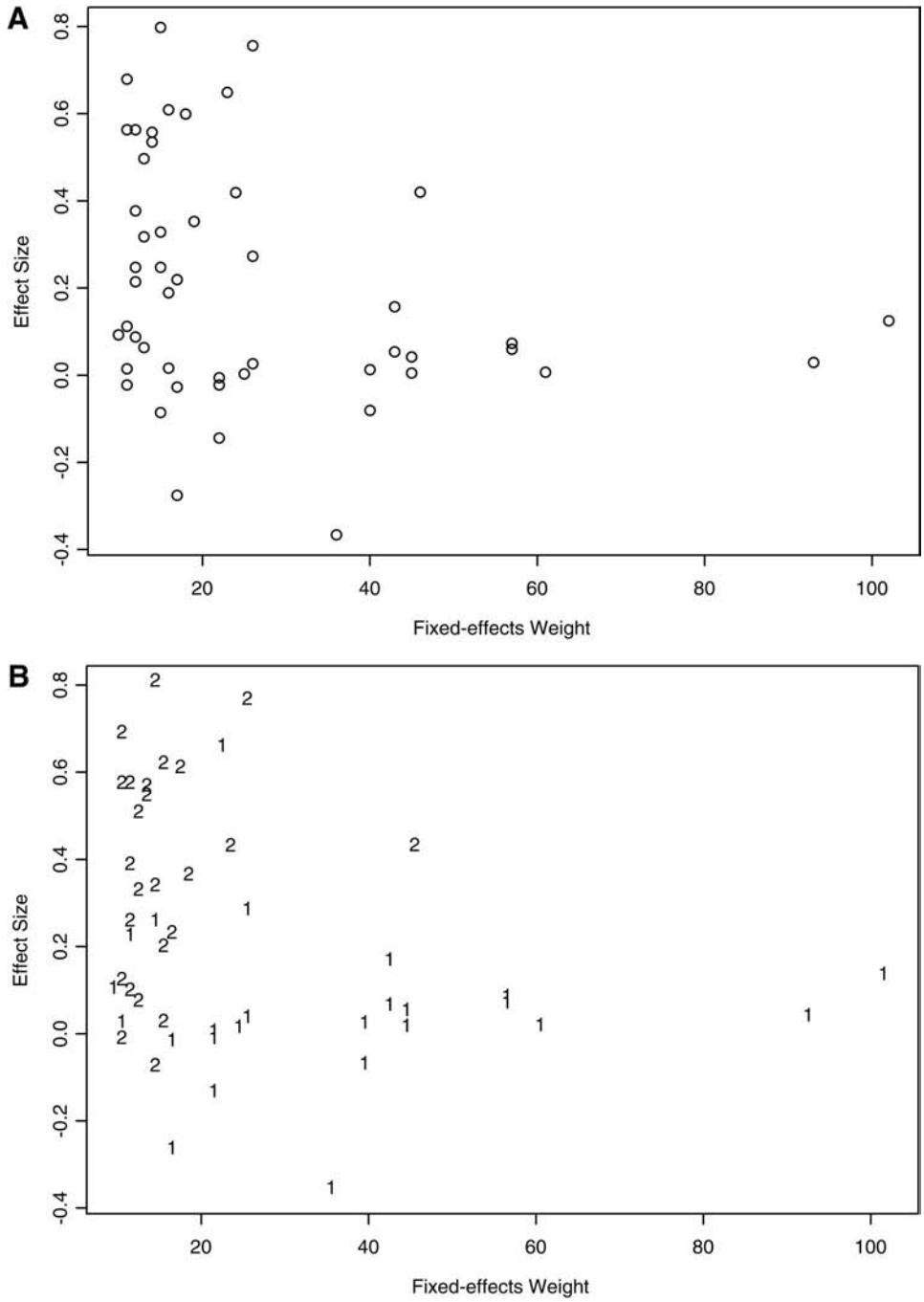

Figure 1. A: Funnel plot of simulated effect sizes against fixed-effects weights, showing funnel-plot asymmetry. B: Funnel plot of simulated effect sizes against fixed-effects weights, showing funnel-plot asymmetry due to mixing two populations, designated 1and 2.

the weight function vary from one model to another. Earlier models used the method of maximum likelihood (ML) to estimate parameters (e.g., Dear & Begg, 1992; Hedges, 1984, 1992; Hedges & Olkin, 1985; Iyengar & Greenhouse, 1988; Patil & Taillie, 1989; Vevea & Hedges, 1995); more recent models have incorporated Bayesian priors and estimation methods such as Gibbs sampling, the Metropolis algorithm, and Monte Carlo integration (e.g., Cleary & Casella, 1997; Givens, Smith, & Tweedie, 1997a; Larose & Dey, 1998; Silliman, 1997).

Weight function models may include statistical tests for publication bias (e.g., chi-square, likelihood ratio, rank cor-

relation) and use a step function for the weight function, which takes on different values at different probability values (Dear & Begg, 1992; Hedges, 1992; Vevea & Hedges, 1995). What they have in common is the ability to detect apparent publication bias and the ability to provide an adjusted estimate that (subject to the assumptions of the model's being correct) offers a better sense of what the true effect is. However, they perform best in situations where there is a great deal of information about the selection process, and that usually means meta-analyses with many (i.e., 100 to several hundred) individual effects.

One recently proposed weight-function model (Copas, 1999; Copas & Li, 1997; Copas & Shi, 2000) assumes that selection depends on both the standard error of the effect-size estimate and the magnitude of the estimated effect. The model represents the probability of selection through two parameters that form a regression-like function, in which an intercept and a slope determine the minimal probability of selection and the rate of increase in the likelihood of selection as the standard error decreases. Although the parameters that represent selection are inestimable, the authors proposed a sensitivity-analysis approach in which plausible values are assessed, and the impact on the meta-analytic estimate can be gauged. The idea of assuming a particular form for a selection function that cannot be estimated, and varying that form as part of a sensitivity analysis, provides the inspiration for the method proposed in this article. We apply the same idea to the Vevea and Hedges (1995) model under circumstances in which that model's weights cannot be estimated because of sparse data.

Sensitivity Analysis Using Weight-Function Models

One limitation of both the trim-and-fill and the weight-function technique is that the result is only as correct as the assumptions that underlie it. Because the publication process is unobserved, theoretical, and likely different for different substantive areas, flexibility is desirable in a correction technique. The model proposed by Vevea and Hedges (1995) probably allows the greatest variety of possible forms for the selection function (subject to the assumption that the selection function is constant over fixed probability value ranges and depends only on the probability value). However, as was noted earlier, this method requires large data sets for stable estimates of the weight function and thus can be difficult to apply to small-scale meta-analyses.

In the current article, we present a modification of the model that can be used in situations in which the number of effects does not support the accurate estimation of the weights. The approach allows the analyst to incorporate linear predictors. This is an important feature, because funnel-plot asymmetry that is indistinguishable from significance-based publication bias can be produced if studies that belong to a distinct population with a larger true effect tend to have small sample sizes. The strategy used in the new method is to adopt the model put forth by Vevea and Hedges (1995) but, rather than estimate the weight function, instead impose a set of fixed weights determined a priori and chosen to represent a specific form and severity of biased selection. The strategy, then, is similar to the Copas (1999) model, in that the problem of an inestimable weight function is circumvented by assessing the impact of various forms of selection as part of a sensitivity analysis. By applying a sequence of such models with various sets of weights that represent different types and severities of selection, the analysts may be able to satisfy themselves that the meta-analytic model is robust to the effects of selection.

In this article, we specify the model, discuss estimation, and present examples of its application to real data sets. Code written for the S-PLUS computing environment is supplied on the *Psychological Methods* Web site for readers who wish to use the method (see http://dx.doi.org/10.1037/1082=989X.10.4.428.supp).

Model and Notation

It is convenient to think of the statistical model in two parts: a model for effect sizes and a model for the selection process. This division parallels the presentation in Vevea and Hedges (1995), but the statistical model differs in two important respects. First, the explanatory model for effect sizes may be a fixed- or a random-effects model. Second, whereas in the previous formulation the model for the selection process involved parameters to be estimated, now these same parameters are *specified* to have particular values, so that the effects on the explanatory model of various possible selection patterns may be assessed.

Model for effect sizes. We use the usual formulation of the meta-analytic explanatory model. Let $T_1, T_2, \ldots, T_n$ denote the effect-size outcomes (e.g., correlations, standardized differences between means, log-odds ratios) from n studies, such that $T_i \sim N(\theta_i, \sigma_i^2)$, where σ_i^2 is the approximately known conditional variance of the effect size. The parameter θ_i may be thought of in different ways, depending upon the exact nature of the model. In the simplest case, $\theta_i = \theta$ for all i, and we have a fixed-effects model, which assumes a single common true effect for all studies. Next, we may allow θ_i to be the outcome of a linear equation involving study characteristics (i.e., in matrix form, $\theta = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is a design matrix of known study characteristics, and $\boldsymbol{\beta}$ is a vector of regression coefficients). The result is a fixed-effects model with covariates, in which we assume a distinct common true effect for each study with an identical combination of known covariates.

Either of those models may be modified by thinking of the parameter θ as representing the mean of a distribution of effects. It is customary (and often reasonable) to assume a normal distribution of effects, so that $T_i \sim N(\theta_i, \sigma_i^2 + \tau^2)$,

where τ^2 is a between-studies variance component. The total variability of T_i now includes a component σ_i^2 that quantifies variation associated with the sampling of persons or other primary units into the study, as well as the variance component τ^2, which quantifies variation that arises from the sampling of the study's population from a distribution of possible populations. When θ_i has a common value $\theta_\bullet$ for all i, this is the simple random-effects model. In that case, $\theta_\bullet$ represents the mean of a distribution of random effects that is normal with variance τ^2. When θ_i is the outcome of a linear model, the result is the random-effects analog of a fixed-effects model with covariates, often called a *mixed-effects model*. In that situation, we think of the intercept in the linear model (β_0) as the mean of a normal distribution of possible population intercepts. As was the case in the simple random-effects model, the intercept in a particular study has variability arising both from the sampling of primary units into the study (σ_i^2) and the variance component (τ^2) that describes variation in the hyperpopulation of intercepts from which the study was sampled. (For further information on mixed- and random-effects models, see Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Raudenbush, 1994).

Model for selection. We saw earlier that models that estimate the effects of selection on the meta-analysis often use a form of weighted distribution theory (e.g., Dear & Begg, 1992; Hedges, 1992; Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995). Given a density f(x) and a nonnegative weight function w(x), the function

$$\frac{\mathrm{w}(x)\mathrm{f}(x)}{\int \mathrm{w}(x)\mathrm{f}(x)dx} \tag{1}$$

will be a density that represents the original density weighted by w(x). For our purposes, f(x) is the normal density appropriate for the model for effect sizes that we intend to use. Vevea and Hedges proposed a step function for w(x) that depended on the magnitudes and variances of the individual effect sizes through their one-tailed probability values.[4] That is, they let

$$\mathrm{w}(p_i) = \begin{cases} \omega_1 \text{ if } 0 < p_i \le a_1; \\ \omega_j \text{ if } a_{j-1} < p_i \le a_j; \text{ and} \\ \omega_k \text{ if } a_{k-1} < p_i \le 1, \end{cases} \tag{2}$$

with $a_1, \ldots, a_k$ representing cutpoints set at prespecified locations. The quantity p_i was the one-tailed probability value of the ith study, approximated by $1 - \Phi(T_i/\sigma_i)$, where $\Phi(z)$ denotes the normal cumulative distribution function evaluated at z. The weights $\omega_2, \ldots, \omega_k$ were regarded as parameters to be estimated. (The first weight was fixed at 1.0 to address an indeterminacy.) The weight function was combined with the normal density appropriate to the chosen model for effect sizes (random- or mixed-effects) through Equation 1. The parameters $\theta_\bullet$ (for a random-effects model) or $\boldsymbol{\beta}$ (the vector of regression coefficients for a mixed-effects model), τ^2, and $\boldsymbol{\omega}$ (the vector of weights) were estimated simultaneously by ML. Most often, no more than 10 distinct weights were used, and the cutpoints were set at psychologically important probability values such as .001, .01, .05, or .50.

The vector of weights is extremely difficult to estimate. Accordingly, although the method works well for large meta-analyses such as the examples presented in Vevea and Hedges (1995) or Vevea, Clements, and Hedges (1993), it is difficult to apply to meta-analyses with fewer than, say, 100 to 200 studies. Typically, one wants at least 10 to 15 studies observed within each probability value interval in order to estimate the weights adequately. Hence, for a meta-analysis with only 30 effects, one would be limited to no more than two probability value intervals, which imposes an unrealistically simple structure on the selection function. We now propose a modification of the approach, in which the weights themselves are set at prespecified values, along with the probability value cutpoints, and the parameters of the model for effect sizes are estimated by ML, conditional on the weights. A disadvantage of this approach is that we may no longer regard the estimates of the effect-size model parameters as good estimates of the parameters as they would exist in the absence of publication bias. Rather, we view them as reasonable estimates if the publication selection function looked exactly as we have specified it. That restriction may seem unreasonable to some readers. We argue, however, that a conventional meta-analysis implicitly does exactly the same thing, with the often implausible specification that the weights are all 1.0. If we estimate the effect-size model under that restriction, and then reestimate it applying a set of weights that represents a plausible pattern of probability-value-based selection, and we find that the estimates are severely attenuated, there is cause for concern. On the other hand, if we apply a variety of reasonable sets of weights representing various different patterns of selection (perhaps guided by the appearance of a funnel plot), and the estimates of the effect-size model parameters are never much affected, we are in a strong position to make the claim that it is unlikely our conclusions are primarily due to publication bias.

If we apply our weight function to the most general model for effects (the mixed-effects model), using Equation

[4] The use of one-tailed probability values in the model in no way implies that one-tailed tests were conducted in the original analysis. A selection function that behaves as if it were based on two-tailed tests can easily be specified in terms of one-tailed probability values, by considering cutpoints such as .950, .990, .995, and so on, in addition to cuts at .05, .01, or .005. In practice, however, when the weights are to be estimated, there will usually be sufficient effect sizes to allow the estimation of weights only in one tail of the distribution.

1, and if we assume that the n individual effects are independent, we get the joint density

$$f(\mathbf{T}|\boldsymbol{\beta},\tau^2;\mathbf{X},\boldsymbol{\sigma},\boldsymbol{\omega}) = \prod_{i=1}^{n} \frac{w(T_i,\sigma_i)/\sqrt{2\pi(\sigma_i^2+\tau^2)} \times \exp(-1/2(T_i-\mathbf{X}_i\boldsymbol{\beta})^2/(\sigma_i^2+\tau^2}{\int_{-\infty}^{\infty} w(T_i,\sigma_i)/\sqrt{2\pi(\sigma_i^2+\tau^2)} \times \exp(-1/2(T_i-\mathbf{X}_i\boldsymbol{\beta})^2/(\sigma_i^2+\tau^2))dT_i}. \quad (3)$$

Other models for effects (i.e., the simple fixed-effects model, the fixed-effects model with covariates, and the simple random-effects model) may be considered as restricted cases of the mixed-effects model. For example, if we set the variance component to zero, we have the fixed-effects model with covariates. If we remove that restriction on the variance component but impose an intercept-only linear model for effects, we have the simple random-effects model. Thus, from Equation 3, we may derive a function proportional to a general log-likelihood for all of the effect-size models,

$$L(\boldsymbol{\beta},\tau^2|\mathbf{T};\boldsymbol{\sigma},\mathbf{X},\boldsymbol{\omega}) = \sum_{i=1}^{n} \frac{\log(w(T_i,\sigma_i)) - 1/2\log(\sigma_i^2+\tau^2) - 1/2(T_i-\mathbf{X}_i\boldsymbol{\beta})^2/(\sigma_i^2+\tau^2)}{\int_{-\infty}^{\infty} \log(w(T_i,\sigma_i)) - 1/2\log(\sigma_i^2+\tau^2) - 1/2(T_i-\mathbf{X}_i\boldsymbol{\beta})^2/(\sigma_i^2+\tau^2)dT_i}. \quad (4)$$

Following Vevea and Hedges (1995), we can simplify this by noting that the integral in the denominator may be treated as a probability-weighted sum over the discrete probability value intervals. Let

$$B_{ij}(\boldsymbol{\beta},\tau^2;\mathbf{X}_i,\sigma_i^2) = \begin{cases} 1-\Phi((b_{i,1}-\mathbf{X}_i\boldsymbol{\beta})/\sqrt{\sigma_i^2+\tau^2}) & \text{if } j=1; \\ \Phi((b_{i,j-1}-\mathbf{X}_i\boldsymbol{\beta})/\sqrt{\sigma_i^2+\tau^2}) - \Phi((b_{ij} & \\ \Phi((b_{i,k-1}-\mathbf{X}_i\boldsymbol{\beta})/\sqrt{\sigma_i^2+\tau^2}) & \\ -\Phi((b_{ij}-\mathbf{X}_i\boldsymbol{\beta})/\sqrt{\sigma_i^2+\tau^2}) & \text{if } 1<j<k; \\ & \text{if } j=k, \end{cases} \quad (5)$$

where b_{ij} is the left endpoint of the range of effect sizes that would fall within the jth probability value interval for the ith study. That is, B_{ij} is the probability that a normally distributed random variable with mean $\mathbf{X}_i\boldsymbol{\beta}$ and variance $\sigma_i^2+\tau^2$ will fall into the jth probability value interval and hence be assigned a weight of ω_j, and $b_{ij} = \sigma_i\Phi^{-1}(a_j)$, where $\Phi^{-1}(p)$ denotes the inverse normal cumulative distribution function evaluated at p. Then Equation 4 may be expressed as

$$L(\boldsymbol{\beta},\tau^2|\mathbf{T};\boldsymbol{\sigma}^2,X,\boldsymbol{\omega}) = -\frac{1}{2}\sum_{i=1}^{n}\frac{(T_i-\mathbf{X}_i\boldsymbol{\beta})^2}{\sigma_i^2+\tau^2} - \frac{1}{2}\sum_{i=1}^{n}\log(\sigma_i^2+\tau^2) - \sum_{i=1}^{n}\log(\sum_{j=1}^{k}\omega_j B_{ij}(\boldsymbol{\beta},\tau^2;\mathbf{X}_i,\sigma_i^2)) \quad (6)$$

This expression is simpler than Vevea and Hedges's (1995) formulation, because terms involving only the weight function and not other parameters are constant and thus may be dropped. Parameter estimates may be obtained by maximizing Equation 6. The details of estimation are presented in the Appendix.

Examples of the Method

The data used for the examples presented here are taken from Cooper and Hedges (1994). In each case, we present a brief description of the data set and give the results of a conventional analysis. We then apply a sequence of four hypothetical weight functions representing different degrees and forms of possible selection. Finally, we describe how one might interpret the sensitivity analysis.

For clarity of presentation, we apply the same four sets of weights to each data set. In practice, one would want to apply a wider range of trial functions, and one might wish to speculate on a likely weight function, given what one observed in a funnel plot. For example, if a funnel plot showed a sudden marked sparseness of negative effects, one might consider dramatically lower weights for probability values above .50. Here, we label the weight functions by the terms *moderate one-tailed selection*, *severe one-tailed selection*, *moderate two-tailed selection* and *severe two-tailed selection*. The specific weights and probability value cutpoints that constitute each of those conditions are presented in Table 1. These are merely examples and should not be regarded as canonical. However, they are based on typical estimated weight functions observed in applications of the Vevea and Hedges (1995) model. Most empirically estimated weight functions resemble the one-tailed selection patterns.

Validity of student ratings. The first illustrative data set (see Table A.3 in Cooper & Hedges, 1994) consists of 20 correlations between student ratings of instructors and subsequent exam performance. The data, taken from Cohen (1983), are limited to correlations based on samples of size 10 or greater. The students were participants in multisection courses in which a common examination was shared across sections. The subject domains of these courses were quite diverse, including psychology, mathematics, engineering, languages, and hard sciences. Figure 2 depicts a funnel plot for the data set. The plot appears to exhibit symmetry, so that it would not cause concern about publication bias. We

Table 1
Specification of Weights in the Sample Analyses

	Probability of observing effect			
p interval	Moderate one-tailed selection	Severe one-tailed selection	Moderate two-tailed selection	Severe two-tailed selection
.000–.005	1.00	1.00	1.00	1.00
.005–.010	.99	.99	.99	.99
.010–.050	.95	.90	.95	.90
.050–.100	.90	.75	.90	.75
.100–.250	.80	.60	.80	.60
.250–.350	.75	.50	.75	.50
.350–.500	.65	.40	.60	.25
.500–.650	.60	.35	.60	.25
.650–.750	.55	.30	.75	.50
.750–.900	.50	.25	.80	.60
.900–.950	.50	.10	.90	.75
.950–.990	.50	.10	.95	.90
.990–.995	.50	.10	.99	.99
.995–1.000	.50	.10	1.00	1.00

applied Fisher's *r*-to-*z* transformation and estimated a simple random-effects model.

The ML estimate of the mean of the population of transformed correlations was 0.38 ($SE = 0.044$), with a variance component of 0.001 ($SE = 0.014$). Transformed back to the original metric of Pearson's product–moment correlation, a 95% confidence interval for the mean correlation would range from .29 to .44, and the point estimate would be .36. From that starting point, we begin our sensitivity analysis.

Table 2 shows the estimates in the Fisher metric with standard errors and shows the comparison to estimates under the four weight-function scenarios. Note that standard errors are not presented for the adjusted estimates. This is to highlight the idea that the adjusted estimates should not be viewed as legitimate estimates in their own right; rather, they would be good estimates if the true selection function were exactly as we have specified, but we have no idea how closely or distantly the actual selection function might resemble the ones we are trying. The researcher conducting the sensitivity analysis, then, should focus not on the particular estimates obtained, but, rather, on the question of how much those estimates move as different weight functions are specified. (If one did require standard errors that were accurate subject to the assumption that the weight function is correctly specified, they could be obtained analytically by inserting corrected parameter values into the Hessian matrix and inverting; however, we do not advocate this practice, as the weight function can never be considered correct.)

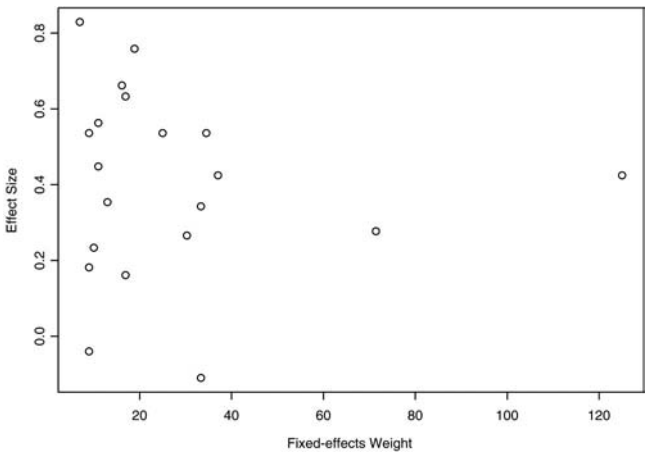

Figure 2. Funnel plot of student rating effects against fixed-effects weights.

When we imposed the weight function denoted moderate one-tailed selection (see Table 1, column 1), the effect was to adjust the estimated mean transformed correlation to 0.35; see the second row of Table 2. This represents an attenuation of the estimate by about 3% of the original value. The estimated variance component increased to

Table 2
Results for Validity of Student Ratings

Selection condition	Transformed correlation	Variance component
No selection	0.38 ($SE = 0.044$)	0.001 ($SE = 0.014$)
Moderate one-tailed	0.35	0.002
Severe one-tailed	0.32	0.005
Moderate two-tailed	0.35	0.002
Severe two-tailed	0.32	0.004

0.002. Thus, if a pattern of selection that tended to favor the publication of significant positive effects had been present (as specified in Table 1), and the data at hand represented the effects that had survived that process, our estimate of the population mean effect would not be substantially altered. If we impose the weight function denoted severe one-tailed selection, in which only a small fraction of large negative correlations survive and fewer than half of any correlations near zero are observed, the picture is quite similar; the reduction in the magnitude of the estimated mean correlation is only about 12% of the original estimate. Similar results hold for two-tailed selection patterns (i.e., in which correlations near zero are less likely to be observed, but significant correlations in either direction are favored). The adjusted estimates appear in the third and fourth rows of Table 2. In no case is the estimated mean transformed correlation moved below 0.32.

All of these outcomes are consistent with the idea that it would be difficult to explain the positive result of the original meta-analysis by arguing that it occurred because of publication bias. In actual practice, we would probably try a larger variety of putative weight functions in our sensitivity analysis; the four we have chosen serve merely to illustrate the procedure. What is clear for this data set, though, is that unless one were to impose a weight function that made correlations with moderate probability values extremely unlikely to survive the publication process, the model would not greatly alter our belief about the true effect magnitude. The result provides support for our initial assessment based on the funnel plot. In a paper that incorporates this meta-analysis, we would report the original, unadjusted values for the correlation and variance component; we would note that under sensitivity analysis using a priori weight functions, the estimate proved quite robust, so that it was unlikely that publication bias is an important counterexplanation for the finding.

Teacher expectancy studies. Next, we consider a slightly more complicated example, in which we use a model with a categorical moderator of effect size (see Table A.2 in Cooper & Hedges, 1994). The data, originally taken from an article by Raudenbush and Bryk (1985), were the basis for an examination of experiments in which it was suggested to teachers that a randomly selected group of students was intellectually gifted. The 19 effect sizes are standardized mean differences on a performance task, comparing the treatment group with a control group of students who were not identified as gifted. It is likely that teachers who have had extensive contact with their students before the supposedly gifted students are identified will have already formed strong impressions of their students' abilities. Accordingly, estimated time of contact prior to the manipulation was considered as a covariate. A scatter plot of effect size against time of contact appears to show that 2 weeks is a critical juncture, and that among teachers who have known their students for less than that period, impressions are more subject to manipulation. We therefore created a categorical variable that identified whether prior contact exceeded 2 weeks' duration. (The data set included 8 effects for which the contact exceeded 2 weeks and 11 short-term effects.) A funnel plot for the data set appears in Figure 3. Initially, the plot appears highly asymmetrical. However, at least part of that asymmetry may be caused by the superimposition of the two types of effect; note that the effects from studies with long-term contact do not appear asymmetrical.

Initially, we fitted a random-effects model that estimated an intercept, a coefficient associated with long-term contact, and a variance component. ML estimates of the model parameters assuming no publication bias resulted in an estimate of 0.0 for the variance component. This is consistent with a homogeneous data set; however, the standard error for the variance component is invalid because the estimate of 0.0 is a border condition. We performed a residual homogeneity test from a conventional fixed-effects moderator analysis; the result was $Q_{17} = 22.81$, $p = .16$. Hence, we assumed homogeneity and changed to a fixed-effects analysis. Table 3 presents results of the fixed-effects analysis. The first two numeric columns of the table list estimates of the intercept and the coefficient for longer contact under each selection scenario. In the last two columns, we see predicted expectancy effects for teachers with no prior contact and for teachers with prior contact.[5] When the moderate one-tailed selection pattern is imposed, the estimated intercept and slope change modestly, resulting in a predicted value for the no-contact condition that is attenuated by about 15%. For cases in which the contact was longer, the change is about 50% of the originally predicted value, although the prediction is still a negligible effect. When we assume severe one-tailed selection, the predicted value for the no-contact condition is about 30% lower than the original prediction, and the model prediction for the contact condition is 133% lower than the original value. The pattern is similar for two-tailed selection.

Although some of these changes might not dramatically alter our perceptions of teacher expectancy, the fact that the estimates can be changed more easily and by larger relative amounts than in the first example is cause for concern. Indeed, under the severe one-tailed condition, the estimates might even be said to overturn our original conclusion, because we have moved from perceiving a small effect for one condition and a negligible effect for the other to a suggestion of small-to-negligible effects (in opposite direc-

[5] Confidence intervals for these predictions could be easily obtained. However, they would require that the analyst take into account the covariance between estimated slope and estimated intercept, a complication that we prefer not to address in the present context.

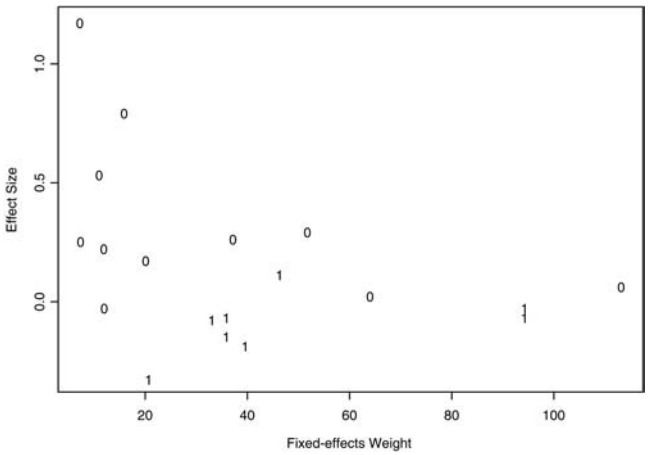

Figure 3. Funnel plot of teacher expectancy effects against fixed-effects weights (0 = short-term contact, 1 = long-term contact).

tions) for both conditions. Under the circumstances, then, although we would probably stand by our initial conclusion that there is some evidence of expectancy effects for the short-duration condition, we would have less confidence in the value of our estimate, because it clearly could have been affected by the problem of publication bias. In a paper that incorporates the meta-analysis, our attention would focus primarily on the no-contact effect. We would mention that a sensitivity analysis using a priori weight functions suggested that the true effect is likely to be somewhat lower (perhaps even by as much as 30%) than the conventional estimate. However, we would note that this most extreme result is probably too extreme, as it predicts a negative effect when there was previous contact, which is not reasonable. Hence, we would argue that there is support for a modest positive effect that is probably somewhat smaller than the original estimate.

Gender differences in conformity. In our final example (Table A.1 in Cooper & Hedges, 1994), we consider a model with a continuous covariate. The effects represent gender differences in studies of conformity that were conducted by informing participants that a normative group had responded to questions in a particular way. In fact, the reference group did not exist. The effects are 10 standardized mean differences in the degree to which male and female participants' responses resembled the fictitious responses. The data were originally taken from a subset of effect sizes considered by Eagly and Carli (1981) and reanalyzed by Becker (1986). The data set presented in the Cooper and Hedges article includes two possible moderators: percentage of male authors and number of items used in the questionnaire. Here, we consider only the second. We estimated conditional variances for the effect sizes under the assumption that the total sample size was equally divided between male and female participants.

A scatter plot of effect size against number of items on the questionnaire suggested a linear relationship in which effect size tends to increase with the length of the questionnaire. Figure 4 depicts a funnel plot for the data set, with the points identified by the number of questions on the questionnaire. The plot suggests a very strong association between sample size and effect size. However, that pattern is confounded by the issue of how many questions were used, so that it is difficult to interpret the plot. We estimated a regression-like model with number of questions as a predictor of effect size. The ML estimate of the intercept was −0.15 (SE = 0.110), the estimated slope was 0.014 (SE = 0.004), and the variance component was 0.002 (SE = 0.022). However, the Q statistic for residual heterogeneity in a conventional fixed-effects analysis was 11.37 on 8 degrees of freedom (p = .18). In the sensitivity analyses that follow, the variance component often is estimated to be zero. In view of the apparent homogeneity of the data set, and of the problematic border condition for the variance component estimates, we again used a fixed-effects approach for the sensitivity analysis.

The number of items in the questionnaires in the various studies ranged from 2 to 45, and the mean number of items was 21.9. In Table 4, we present estimates of the model parameters under each selection condition, as well as predicted conformity effects for scales with the minimum (2), the maximum (45), and the average (21.9) number of items. The only selection condition that had any appreciable effect on the parameter estimates of the model predictions was the severe one-tailed condition, where the predicted negative effect for 2-item scales was

Table 3
Results for Teacher Expectancy

Selection condition	Estimated intercept	Estimated difference for longer contact	Predicted effect for no prior contact	Predicted effect for prior contact
No selection	0.20 (SE = 0.053)	−0.26 (SE = 0.079)	0.20	−0.06
Moderate one-tailed	0.17	−0.26	0.17	−0.09
Severe one-tailed	0.14	−0.28	0.14	−0.14
Moderate two-tailed	0.18	−0.24	0.18	−0.06
Severe two-tailed	0.16	−0.20	0.18	−0.04

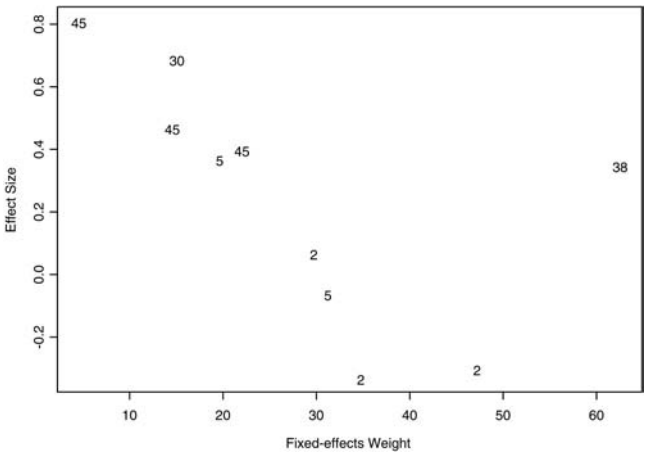

Figure 4. Funnel plot of gender difference effects against fixed-effects weights; plotting symbols are the number of items on the scale.

almost twice as great as the original estimate. Even in this condition, though, the predicted conformity effects for average-length and long scales did not differ appreciably from the unadjusted estimates. Hence, the analysis indicates that, despite the suggestive funnel plot, there may not be great cause for concern about the validity of the analysis; estimated effects were malleable under different possible selection conditions only for the 2-item scales, and these are probably not the ones we would want to focus on when assessing conformity. In a paper that incorporated this meta-analysis, then, we would note that a sensitivity analysis using a priori weight functions suggested that the effects were quite robust for the studies that used larger scales. We would conclude, then, that publication bias is unlikely to represent a serious threat to validity here and would focus our discussion on the unadjusted estimates.

Discussion

Whenever a statistical conclusion is reached, whether it be a hypothesis test or a statement that a parameter of interest is thought to have some particular value, the skeptic may question the conclusion by questioning the assumptions that underlie it. If the statistical conclusion is presented in the context of a meta-analysis, one assumption that will always be required for the result to be defensible is that the effect-size estimates on which the conclusion is based represent either a near-perfect census of, or an unbiased sample from, the universe of effect sizes that exist. Hence, it will usually be appropriate for the meta-analysts to present a set of arguments against the counterexplanation that has been the subject of this article, namely, that the process of submission and publication systematically leads to violations of that assumption.

In the best examples of quantitative synthesis, the authors describe extensive measures that they have undertaken to seek out the fugitive literature. That is indeed an essential part of sound meta-analysis; yet it would be difficult to argue that any effect-size retrieval effort, no matter how exhaustive, could unlock all of the file drawers. Thus, even in the presence of such efforts, the problem of publication bias cannot be entirely discounted. The author of the meta-analysis, then, is faced with a logically impossible task: to show that publication bias is not a problem for the particular data set at hand.

We describe the task as logically impossible because it amounts, in essence, to an attempt at confirming a null hypothesis. The conflict is easily understood by considering a similar issue in a more familiar situation. In the context of a two-sample *t* test, one important assumption is that the two groups being compared represent draws from populations with equal variance. An obvious way to investigate that assumption is to conduct an auxiliary test of the null hypothesis that the variances are equal. However, the desired outcome of that test is a failure to reject the null. If the sample size is small, it is quite likely that the test will fail to detect unequal variances, not because the variances are equal, but because the test lacks power. Upon reflection, we see that the test will necessarily lack power in precisely the circumstances in which its outcome is most important for the validity of the *t* test that is our real interest: when the sample size is small. Hence, the responsible statistician will use a battery of approaches to investigate the possibility of unequal variances, including not only the auxiliary test but

Table 4
Results for Gender Differences in Conformity

Selection condition	Estimated intercept	Estimated coefficient for number of items	Predicted effect for 2 items	Predicted effect for 45 items	Predicted effect for 21.9 items
No selection	−0.15 (*SE* = 0.110)	0.014 (*SE* = 0.004)	−0.12	0.48	0.16
Moderate one-tailed	−0.15	0.015	−0.12	0.53	0.18
Severe one-tailed	−0.25	0.016	−0.22	0.47	0.10
Moderate two-tailed	−0.13	0.014	−0.10	0.50	0.18
Severe two-tailed	−0.11	0.013	−0.08	0.48	0.17

also informal comparisons of the magnitudes of the sample standard deviations and, perhaps, graphical comparisons of variability in the two samples. If those approaches all seem to point to the conclusion that heterogeneity is not an issue, then the researcher is in a strong position to refute the proposition that the statistical conclusion has been biased by its presence.

We advocate a similar approach to the problem of publication bias in meta-analysis. We have described a variety of existing tools to address the problem and have presented one new approach. None of these methods should be viewed as sufficient in its own right. The funnel plot can be difficult to interpret. On the one hand, quite severe examples of publication bias may be difficult to observe in a funnel plot. On the other hand, the presence of a subgroup of effects that are drawn from a different population from the rest and also have systematically different sample sizes may create the appearance of publication bias in a funnel plot for which it is truly not a problem. The fail-safe *n* in its several forms, while not difficult to apply, often addresses a question that is not quite the question of interest and should rarely be used. The trim-and-fill method is easy to apply even to small data sets and has the advantage of providing a direct estimate of effect magnitude adjusted for bias. However, its assumptions are strong; in particular, the assumption that missing effects are deterministically ordered, with the most extreme unexpected values necessarily missing first, suggests that its estimate might appropriately be viewed as a lower bound for the effect magnitude. Moreover, it does not address the problem of funnel-plot asymmetry that is due in part to moderating variables and has been seen to be biased in the presence of heterogeneity. Methods that use weighted distribution theory make their own set of strong assumptions but in at least one model have the advantage of being able to tease out the effects of moderating variables from the effects of publication bias. However, the weights in these models are difficult to estimate, so that they are appropriate only for large data sets.

The method presented in this article addresses that difficulty by eliminating the need to estimate the weight function. In so doing, the model gives up the ability to present an adjusted estimate that represents a best guess at the true effect magnitude. Nevertheless, it represents an additional tool that is useful in the attempt to refute publication bias as a counterexplanation for meta-analytic findings.

In this article, we saw three examples of sensitivity analysis using the new approach. For the first of these (validity of student ratings) and the third (gender differences in conformity), although it was possible to apply a hypothetical weight function that would change the effect-size estimate or the estimates of coefficients in the linear model, the changes were not dramatic. Regardless of which weight function was applied, the conclusion remained essentially the same: The validity of student ratings was somewhat greater than .3, and there did appear to be a modest gender difference in conformity behavior, particularly when the study used a survey measure with an adequate number of items. By contrast, in the second example (teacher expectancy effects), the outcome of greatest interest (effect magnitude when the teacher was not well acquainted with the students at the onset of the study) was more malleable as different hypothetical weight functions were applied. Moreover, one scenario actually changed the effect-size estimate for the less interesting case from negligible to a modest negative outcome. The authors of studies reporting the first and third examples, then, would be in a strong position to argue against the phenomenon of publication bias as a primary explanation for their findings, even though the funnel plots may have suggested a problem. A hypothetical author of the second example should probably acknowledge the possibility that publication bias may be an issue for the analysis. Not only does the estimated effect size appear to be vulnerable but this occurs in the context of a funnel plot that strongly suggests a systematic relation between sample size and effect size.

It is our hope that it will become increasingly common practice to address publication bias as a threat to validity in meta-analysis by performing sensitivity analysis with a variety of techniques. The method described here adds to the arsenal of available techniques one that fills a void: It addresses the problem of publication bias in small data sets in which moderating variables may contribute to the appearance that bias is present.

References

Ashworth, S. D., Osburn, H. G., Callender, J. C., & Boyle, K. A. (1992). The effects of unrepresented studies on the robustness of validity generalization results. *Personnel Psychology, 45,* 341–361.

Becker, B. J. (1986). Influence again. In J. S. Hyde & M. L. Linn (Eds.), *The psychology of gender: Progress through meta-analysis* (pp. 178–209). Baltimore: Johns Hopkins University Press.

Becker, B. J. (in press). Failsafe *N* of file-drawer number. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 111–126). Chichester, England: Wiley.

Begg, C. B. (1985). A measure to aid in the interpretation of published clinical trials. *Statistics in Medicine, 4,* 1–9.

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *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. *Journal of the Royal Statistical Society, Series A, 151,* 1–27.

Begg, C. B., & Berlin, J. A. (1989). Publication bias and dissemination of clinical research. *Journal of the National Cancer Institute, 81,* 107–115.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088–1101.

Berlin, J. A., Begg, C. B., & Louis, T. A. (1989). An assessment of publication bias using a sample of published clinical trials. *Journal of the American Statistical Association, 84,* 381–392.

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

Chalmers, I. (1990). Underreporting research is scientific misconduct. *Journal of the American Medical Association, 263,* 1405–1408.

Chalmers, I., Adams, M., Dickersin, K., Hetherington, J., Tarnow-Mordi, W., Meinert, C., et al. (1990). A cohort study of summary reports of controlled trials. *Journal of the American Medical Association, 263,* 1401–1404.

Chan, S. S., Sacks, H. S., & Chalmers, T. C. (1982). The epidemiology of unpublished randomized control trials. *Clinical Research, 30,* 234A.

Cleary, R. J., & Casella, G. (1997). An application of Gibbs sampling to estimation in meta-analysis: Accounting for publication bias. *Journal of Educational and Behavioral Statistics, 22,* 141–154.

Cohen, P. A. (1983). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51,* 281–309.

Cooper, H. M., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2,* 447–452.

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

Copas, J. B. (1999). What works? Selectivity models and meta-analysis. *Journal of the Royal Statistical Society, Series A, 161,* 95–105.

Copas, J. B., & Li, H. G. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society, Series B, 59,* 55–95.

Copas, J. B., & Shi, J. Q. (2000). Meta analysis, funnel plots and sensitivity analysis. *Biostatistics, 1,* 247–262.

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.

Dawid, A. P., & Dickey, J. M. (1977). Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association, 72,* 845–850.

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

Denton, F. T. (1987). The power function of a published hypothesis test. *Economics Letters, 25,* 101–104.

Denton, F. T. (1990). The effects of publication selection on test probabilities and estimator distributions. *Risk Analysis, 10,* 131–136.

Dickersin, K., Chan, S. S., Chalmers, T. C., Sacks, H. S., & Smith, H. (1987). Publication bias in randomized control trials. *Controlled Clinical Trials, 8,* 343–353.

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

Duval, S., & Tweedie, R. (2000a). A nonparametric trim and fill method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89–99.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463.

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

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

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal, 315,* 629–634.

Egger, M., Zellweger-Zähner, T., Schneider, M., Junker, C., Lengeler, C., & Antes, G. (1997). Language bias in randomized controlled trials published in English and German. *Lancet, 350,* 326–329.

Givens, G. H., Smith, D. D., & Tweedie, R. L. (1997a). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science, 12,* 221–240.

Givens, G. H., Smith, D. D., & Tweedie, R. L. (1997b). Rejoinder. *Statistical Science, 12,* 247–250.

Gleser, L. J., & Olkin, I. (1996). Models for estimating the number of unpublished studies. *Statistics in Medicine, 15,* 2493–2507.

Grégoire, G., Derderian, F., & Le Lorier, J. (1995). Selecting the language of the publications included in the meta-analysis—Is there a Tower-of-Babel bias? *Journal of Clinical Epidemiology, 48,* 159–163.

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

Greenwald, S., & Russell, R. L. (1991). Assessing rationales for inclusiveness in meta-analytic samples. *Psychotherapy Research, 1,* 17–24.

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 bias selection effects in meta-analysis. *Statistical Science, 7,* 246–255.

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects meta-analysis. *Psychological Methods, 3,* 486–504.

Higginson, J. (1987). Publication of "negative" epidemiologic studies. *Journal of Chronic Disease, 40,* 371–372.

Hubbard, R., & Armstrong, J. S. (1997). Publication bias against null results. *Psychological Reports, 80,* 337–338.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. ransom effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8,* 275–292.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science, 3,* 109–117.

Koren, G., Graham, K., Shear, H., & Einarson, T. (1989). Bias against the null hypothesis: The reproductive hazards of cocaine. *The Lancet, 2,* 1440–1442.

Koren, G., & Klein, N. (1991). Bias against negative studies in newspaper reports of medical research. *Journal of the American Medical Association, 266,* 1824–1826.

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

Larose, D. T., & Dey, D. K. (1998). Modeling publication bias using weighted distributions in a Bayesian framework. *Computational Statistics and Data Analysis, 26,* 279–302.

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

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48,* 1181–1209.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1,* 161–175.

Misakian, A. L., & Bero, L. A. (1998). Publication bias and research on passive smoking: Comparison of published and unpublished studies. *Journal of the American Medical Association, 280,* 250–253.

Newcombe, R. G. (1987). Towards a reduction of publication bias. *British Medical Journal, 295,* 656–659.

Nunnally, J. (1960). The place of statistics in psychology. *Education and Psychological Measurement, 20,* 641–650.

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

Patil, G. P., & Taillie, C. (1989). Probing encountered data, meta-analysis and weighted distribution methods. In Y. Dodge (Ed.). *Statistical data analysis and inferences* (pp. 317–345). Amsterdam: Elsevier Science.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–332). New York: Russell Sage Foundation.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 75–98.

Rennie, D., & Flanagin, A. (1992). Publication bias. *Journal of the American Medical Association, 267,* 411–413.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185–193.

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

Rotton, J., Foos, P. W., Vanmeek, L., & Levitt, M. (1995). Publication practices and the file drawer problem: A survey of published authors. *Journal of Social Behavior and Personality, 10,* 1–13.

Scherer, R. W., Dickersin, K., & Langenberg, P. (1994). Full publication of results initially presented in abstracts: A meta-analysis. *Journal of the American Medical Association, 272,* 158–162.

Shadish, W. R., Doherty, M., & Montgomery, L. M. (1989). How many studies are in the file drawer? An estimate from the family/marital therapy literature. *Clinical Psychology Review, 9,* 589–603.

Silliman, N. P. (1997). Nonparametric classes of weight functions to model publication bias. *Biometrika, 84,* 909–918.

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

Simes, R. J. (1987). Confronting publication bias: A cohort design for meta-analysis. *Statistics in Medicine, 6,* 11–30.

Smart, R. G. (1964). The importance of negative results in psychological research. *Canadian Psychologist, 5,* 225–232.

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

Sohn, D. (1996). Publication bias and the evaluation of psychotherapy efficacy in reviews of the research literature. *Clinical Psychology Review, 16,* 147–156.

Sommer, B. (1987). The file drawer effect and publication rates in menstrual cycle research. *Psychology of Women Quarterly, 11,* 233–242.

Sterling, T. D. (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.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 49,* 108–113.

Stern, J. M., & Simes, R. J. (1997). Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal, 315,* 640–645.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal, 320,* 1574–1577.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22,* 2113–2126.

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.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60,* 419–435.

Wang, M. C., & Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods, 3,* 46–54.

Weber, E. J., Callaham, M. L., Wears, R. L., Barton, C., & Young,

G. (1998). Unpublished research from a medical specialty meeting: Why investigators fail to publish. *Journal of the American Medical Association, 280,* 257–259.

Wolff, W. M. (1973). Publication problems in psychology and an explicit evaluation schema for manuscripts. *American Psychologist, 28,* 257–261.

Appendix

Estimation

Consider the parameter vector $\xi = (\boldsymbol{\beta},\tau^2)$. Likelihood equations for estimation of the parameters may be developed from Equation 6 in the main text by taking the first derivatives with respect to the components of ξ. Maximum-likelihood estimates are then obtained by solving the system of equations

$$\frac{\partial \mathrm{L}(\xi|\mathrm{T};\, \sigma^2,\, \mathbf{X},\, \boldsymbol{\omega})}{\partial \xi} = \mathbf{0}. \tag{A1}$$

When the parameter is a component of β, the derivative is

$$\frac{\partial \mathrm{L}}{\partial \beta_j} = \sum_{i=1}^{n} X_{ij} \frac{T_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma_i^2 + \tau^2} \sum_{i=1}^{n} \frac{\sum_{m=1}^{k} \omega_m \frac{\partial \mathrm{B}_{im}}{\partial \beta_j}}{\sum_{m=1}^{k} \omega_m \mathrm{B}_{\mathrm{im}}}, \tag{A2}$$

with

$$\frac{\partial \mathrm{B}_{im}}{\partial \beta_j} = \begin{cases} \dfrac{X_{ij}\varphi\left(\dfrac{b_{i1} - \mathbf{X}_i\boldsymbol{\beta}}{\sqrt{\sigma_i^2 + \tau^2}}\right)}{\sqrt{\sigma_i^2 + \tau^2}} & \text{if } m = 1; \\ \dfrac{X_{ij}}{\sqrt{\sigma_i^2 + \tau^2}}\left(\varphi\left(\dfrac{b_{im} - \mathbf{X}_i\boldsymbol{\beta}}{\sqrt{\sigma_i^2 + \tau^2}}\right) - \varphi\left(\dfrac{b_{i,m-1} - \mathbf{X}_i\boldsymbol{\beta}}{\sqrt{\sigma_i^2 + \tau^2}}\right)\right) & \text{if } l < m < k;, \\ \dfrac{-X_{ij}\varphi\left(\dfrac{b_{i,k-1} - \mathbf{X}_i\boldsymbol{\beta}}{\sqrt{\sigma_i^2 + \tau^2}}\right)}{\sqrt{\sigma_i^2 + \tau^2}} & \text{if } m = k, \end{cases} \tag{A3}$$

where $\varphi(z)$ denotes the standard normal density evaluated at z. When the parameter is the variance component, the derivative is

$$\frac{\partial \mathrm{L}}{\partial \tau^2} = \frac{1}{2}\sum_{i=1}^{n}\left(\left(\frac{T_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma_i^2 + \tau^2}\right)^2 - \frac{1}{\sigma_i^2 + \tau^2}\right) - \sum_{i=1}^{n} \frac{\sum_{m=1}^{k} \omega_m \frac{\partial \mathrm{B}_{im}}{\partial \tau^2}}{\sum_{m=1}^{k} \omega_m \mathrm{B}_{im}}, \tag{A4}$$

where

$$\frac{\partial \mathrm{B}_{\mathrm{im}}}{\partial \tau^2}=\begin{cases} \dfrac{\mathrm{b}_{\mathrm{i1}}-\mathbf{X}_i\boldsymbol{\beta}}{2(\sigma_i^2+\tau^2)^{\frac{3}{2}}}\varphi\left(\dfrac{b_{i1}-\mathbf{X}_i\boldsymbol{\beta}}{\sqrt{\sigma_i^2+\tau^2}}\right) & \text{if } m=1 \\ \dfrac{b_{im}-\mathbf{X}_i\boldsymbol{\beta}}{2(\sigma_i^2+\tau^2)^{\frac{3}{2}}}\varphi\left(\dfrac{b_{im}-\mathbf{X}_i\boldsymbol{\beta}}{\sqrt{\sigma_i^2+\tau^2}}\right)-\dfrac{b_{i,m-1}-\mathbf{X}_i\boldsymbol{\beta}}{2(\sigma_i^2+\tau^2)^{\frac{3}{2}}}\varphi\left(\dfrac{b_{i,m-1}-\mathbf{X}_i\boldsymbol{\beta}}{\sqrt{\sigma_i^2+\tau^2}}\right) & \text{if } l<m<k; \\ \dfrac{b_{i,k-1}-\mathbf{X}_i\boldsymbol{\beta}}{2(\sigma_i^2+\tau^2)^{\frac{3}{2}}}\varphi\left(\dfrac{b_{i,k-1}-\mathbf{X}_i\boldsymbol{\beta}}{\sqrt{\sigma_i^2+\tau^2}}\right) & \text{if } m=k. \end{cases} \tag{A5}$$

The likelihood equations may be solved using the Newton–Raphson iterative method. The forms of the second and partial cross-derivatives are not presented here, because they follow in a straightforward fashion from the forms given in the Appendix of Vevea and Hedges (1995). They are of less interest in the present approach, as we do not advocate using the Hessian matrix to derive standard errors. Standard errors would not be meaningful in the method we propose, because the parameter estimates that arise from the current method should be regarded as aids to sensitivity analysis and not as reasonable estimates of true parameter values adjusted for the effects of publication bias.

The components of the likelihood equations presented above are the most general forms. For fixed-effects models, the variance component τ^2 is set to zero, and derivatives with respect to τ^2 are not used. For the simple random-effects model, the matrix $\mathbf{X}$ consists of a vector of ones. For the simple fixed-effects model without covariates, the matrix $\mathbf{X}$ consists of a vector of ones, and the variance component is set to zero.

An S-PLUS Implementation

The likelihood equations may also be solved using quasi-Newton methods that numerically approximate some or all of the derivatives. We advocate that approach here, partly because it allows us to provide platform-independent code that will allow the user to implement the method in S-PLUS. In addition, using such available software simplifies the problem of imposing constraints, such as the reasonable requirement that the variance component be nonnegative. Computer code for estimating the model using S-PLUS is available at http://dx.doi.org/10.1037/1082=989X.10.4.428.supp

Received February 25, 2003
Revision received March 16, 2005
Accepted June 20, 2005 ■