Psychological Methods
2009, Vol. 14, No. 3, 225–238

1082-989X/09/$12.00 DOI: 10.1037/a0016619

Meta-Analytic Interval Estimation for Standardized and Unstandardized Mean Differences

Douglas G. Bonett
Iowa State University

The fixed-effects (FE) meta-analytic confidence intervals for unstandardized and standardized mean differences are based on an unrealistic assumption of effect-size homogeneity and perform poorly when this assumption is violated. The random-effects (RE) meta-analytic confidence intervals are based on an unrealistic assumption that the selected studies represent a random sample from a large superpopulation of studies. The RE approach cannot be justified in typical meta-analysis applications in which studies are nonrandomly selected. New FE meta-analytic confidence intervals for unstandardized and standardized mean differences are proposed that are easy to compute and perform properly under effect-size heterogeneity and nonrandomly selected studies. The proposed meta-analytic confidence intervals may be used to combine unstandardized or standardized mean differences from studies having either independent samples or dependent samples and may also be used to integrate results from previous studies into a new study. An alternative approach to assessing effect-size heterogeneity is presented.

Keywords: effect size, fixed effects, random effects, between-subjects designs, within-subjects designs

Population means μ_1 and μ_2 may be estimated from a study having two independent samples or two dependent samples. In applications in which the metric of the dependent variable is well understood, a confidence internal for $\mu_1 - \mu_2$ will convey important information. In applications in which the metric of the dependent variable is not well understood, it may be more appropriate to use a standardized difference between two population means of the form $(\mu_1 - \mu_2)/[(\sigma_1^2 + \sigma_2^2)/2]^{1/2}$, where σ_1^2 and σ_2^2 are population variances (Cohen, 1988, p. 44). Both measures of effect size are appropriate in studies having either independent samples or dependent samples.

Confidence intervals for $\mu_1 - \mu_2$, for designs with independent or dependent samples, can be found in many introductory statistics texts. For designs with independent samples in which the homoscedasticity assumption (i.e., $\sigma_1^2 = \sigma_2^2$) cannot be justified, a confidence interval for $\mu_1 - \mu_2$ is obtained with a Satterthwaite adjustment to the degrees of freedom (see, e.g., Snedecor & Cochran, 1980, p. 97). Confidence intervals for $(\mu_1 - \mu_2)/[(\sigma_1^2 + \sigma_2^2)/2]^{1/2}$ that do not assume homoscedasticity are given in Bonett (2008a).

Estimates of effect size for a particular dependent variable are often assessed in several different studies. In study *i*, assume that a random sample is obtained from a specific study population and statistical inference is used to make some statement about the unknown value of $\varphi_i = (\mu_{i1} - \mu_{i2})$ or $\delta_i = (\mu_{i1} - \mu_{i2})/[(\sigma_{i1}^2 + \sigma_{i2}^2)/2]^{1/2}$. A more precise estimate of the effect size might be obtained by combining effect-size estimates from two or more studies, a practice referred to as meta-analysis (Glass, 1976). Estimates with greater precision have smaller standard errors and produce narrower confidence intervals. The main purpose of averaging effect-size estimates from several studies is to obtain an estimate of the average effect size that is more precise than an effect-size estimate from a single study. The confidence interval for the average effect-size value will often be considerably narrower, and hence more informative, than the effect-size confidence interval obtained from a single study. Cohen (1994, p. 1002) has suggested that one reason why researchers are reluctant to report confidence interval results for measures of effect size is that the confidence intervals are often "embarrassingly large." Meta-analysis is one way to obtain more narrow confidence intervals. An increase in external validity is an added benefit of averaging effect-size estimates from multiple studies.

Bond, Wiitala, and Richard (2003) described a fixed-effects (FE) meta-analytic confidence interval for an average unstandardized effect size, $\varphi = (\varphi_1 + \varphi_2 + \cdots + \varphi_m)/m$. Hedges and Vevea (1998) described an FE meta-analytic confidence interval for an average standardized

Correspondence concerning this article should be addressed to Douglas G. Bonett, Department of Statistics, Iowa State University, Ames, IA 50011. E-mail: dgbonett@iastate.edu

effect size, $\delta = (\delta_1 + \delta_2 + \cdots + \delta_m)/m$. The FE methods assume that the m studies have been deliberately selected and that statistical inference applies only to the m study populations represented in the m studies. The average effect size, φ or δ, is a meaningful and interesting parameter to estimate if the population effect sizes are not too disparate across the m study population. The m sample sizes in an FE meta-analysis are typically unequal, and it can be shown that the classical weighted average method of estimating φ or δ can be severely biased when the m population effect sizes are not identical (see Appendix). For this reason, Hunter and Schmidt (2000) and the National Research Council (1992) have recommended that the classic FE methods no longer be used.

Random-effects (RE) meta-analysis methods have been proposed in an attempt to accommodate effect-size heterogeneity. Bond et al. (2003) described an RE meta-analysis method for analyzing unstandardized mean differences. Hedges and Vevea (1998) described an RE meta-analysis method for analyzing standardized mean differences. RE meta-analysis is fundamentally different from FE meta-analysis. In an RE meta-analysis, the researcher must clearly define a very large superpopulation of N study populations from which m studies have been randomly sampled. The set of unstandardized population effect sizes in the superpopulation is $\varphi_1, \varphi_2, \ldots, \varphi_N$, and the set of standardized population effect sizes in the superpopulation is $\delta_1, \delta_2, \ldots, \delta_N$. The N population effect sizes are not assumed to be equal, and the researcher will want to obtain interval estimates of both the mean and the standard deviation of the N population effect sizes. The traditional interval estimation methods for the mean and standard deviation of the N population effect sizes assume that the m studies are a random sample from a superpopulation of N studies and that the N population effect sizes follow an approximate normal distribution. To characterize the degree of effect-size heterogeneity accurately, a narrow confidence interval for the effect-size standard deviation is required, and a large value of m may be needed to achieve an acceptably narrow confidence interval. Furthermore, the traditional confidence intervals for the effect-size standard deviation perform poorly under a nonnormal superpopulation of effect sizes (Viechtbauer, 2007).

The critical random sampling assumption of the RE methods will almost never be satisfied in practice (Hedges & Vevea, 1998), and Schulze (2004, p. 41) warned that the random sampling assumption "is not feasible in practice and may represent a critical point for the application of RE models." The random sampling assumption in the RE methods cannot be taken lightly. Without random sampling, statistical inference cannot be used to generalize from the m study populations to the superpopulation. Some researchers might argue that the m study populations could be considered a random sample from some imaginary superpopulation. However, if that were the case, then statistical inference to the imaginary superpopulation would have limited scientific value, because the researcher may not be able to clearly describe the characteristics of the superpopulation for which the statistical results apply. A detailed description of the population to which results apply is an essential component of any scientific study.

Raudenbush (1994, p. 304) has argued that the unrealistic random sample assumption of RE meta-analysis methods is not required from a Bayesian view that "avoids the specification of any sampling mechanism as a justification of the random effects model." From a Bayesian perspective, the mean and standard deviation of the superpopulation of population effect sizes are viewed as parameters of a prior distribution. The weakness of this argument becomes clear when one tries to describe the prior distribution in an effort to clearly specify the target of statistical inference. Conceptually, this Bayesian prior distribution is no different from the imaginary superpopulation distribution of population effect sizes that a meta-analyst might conjure up when using an RE method to analyze a convenience sample, rather than a true random sample, of m studies.

The above review of FE and RE meta-analysis methods is disheartening and suggests that the multitude of studies that apply the standard meta-analysis methods each year (Hunter & Schmidt, 2004, p. 25) may be producing misleading results. Bonett (2008b) recently proposed an alternative FE meta-analysis method, using an unweighted average rather than a weighted average, for combining Pearson, Spearman, or partial correlation coefficients across multiple studies. The method of unweighted averages for correlations is easy to compute and performs properly under the typical conditions of effect-size heterogeneity and nonrandomly selected studies. The method of unweighted averages is applied here for combining unstandardized mean differences or standardized mean differences from multiple studies. The proposed FE method is general and may be applied to studies that use independent samples or dependent samples. Unlike the methods of Hedges and Vevea (1998) and Bond et al. (2003), the proposed FE method does not require the population variances to be equal within or across the m study population.

Proposed Confidence Intervals

Let $\hat{\varphi}_i = \hat{\mu}_{i1} - \hat{\mu}_{i2}$ denote an estimator of φ_i and $\hat{\delta}_i = \hat{\varphi}_i/[(\hat{\sigma}_{i1}^2 + \hat{\sigma}_{i2}^2)/2]^{1/2}$ denote an estimator of δ_i obtained from study i ($i = 1$ to m), where $\hat{\mu}_{ij}$ is a sample mean and $\hat{\sigma}_{ij}^2$ is an unbiased sample variance for treatment j ($j = 1, 2$). The following point estimator of $\varphi = m^{-1}\sum_{i=1}^{m}\varphi_i$ is proposed,

$$\bar{\varphi} = m^{-1}\sum_{i=1}^{m}\hat{\varphi}_i, \quad (1)$$

and the following point estimator of $\delta = m^{-1}\sum_{i=1}^{m}\delta_i$ is proposed,

$$\bar{\delta} = m^{-1}\sum_{i=1}^{m} b_i\hat{\delta}_i, \quad (2)$$

where b_i is an approximate bias adjustment. Equations 1 and 2 are both unweighted averages and belong to the class of *analog estimates* (Goldberger, 1991, p. 117).

When $\hat{\delta}_i$ is estimated in a study with two independent samples, we set $b_i = 1 - 3/[4(n_{1i} + n_{2i}) - 9]$, which was originally proposed by Hedges (1981) for a similar measure of effect size that assumes equal population variances. A preliminary investigation found that the Hedges bias adjustment also reduces the bias of $\hat{\delta}_i$. When $\hat{\delta}_i$ is estimated in a study with two dependent samples, there is no previous work to suggest an appropriate biased adjustment. In a preliminary investigation, it was found that that the bias of $\hat{\delta}_i$ in dependent samples depends on the magnitude of the Pearson correlation between the n_i paired observations and that setting $b_i = [(n_i - 2)/(n_i - 1)]^{1/2}$ reduced the bias for any correlation value. Note that the confidence intervals for δ proposed by Bonett (2008a) for independent or dependent samples did not employ bias adjustments because the bias of a single estimator of δ is negligible unless the sample size is very small. However, in a meta-analysis in which many estimators of δ_i are combined or compared and the bias of $\hat{\delta}_i$ may vary considerably across studies because of unequal sample sizes, it is best to reduce the bias of each estimator.

An estimate of the variance of $\hat{\varphi}_i$ is

$$\text{var}(\hat{\varphi}_i) = \hat{\sigma}_{i1}^2/n_{i1} + \hat{\sigma}_{i2}^2/n_{i2} \quad (3)$$

for studies that use independent samples of sizes n_{i1} and n_{i2} within study i (see, e.g., Snedecor & Cochran, 1980, p. 96). For studies that use dependent samples, an estimate of the variance of $\hat{\varphi}_i$ is

$$\text{var}(\hat{\varphi}_i) = \hat{\sigma}_{di}^2/n_i, \quad (4)$$

where $\hat{\sigma}_{di}^2$ is an estimated variance of the n_i difference scores in study i. Equations 3 and 4 do not assume equal population variances within studies. Note that $\hat{\sigma}_d^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 = n\hat{\varphi}^2/t^2$, where t is the paired-samples t statistic and $\hat{\rho}$ is the Pearson correlation of the paired observations. In meta-analytic research, the analyst may have access to $\hat{\sigma}_d^2$, $n\hat{\varphi}^2/t^2$, or the sample correlation and variances depending on how the results are reported in each study.

An estimate of the variance of $\hat{\delta}_i$ is

$$\text{var}(\hat{\delta}_i) = [\hat{\delta}_i^2(\hat{\sigma}_{i1}^4/df_{i1} + \hat{\sigma}_{i2}^4/df_{i2})/8\hat{\sigma}_i^4 + (\hat{\sigma}_{i1}^2/df_{i1} + \hat{\sigma}_{i2}^2/df_{i2})/\hat{\sigma}_i^2] \quad (5)$$

for studies that use independent samples of sizes n_{i1} and n_{i2} within study i, where $df_{ij} = n_{ij} - 1$ and $\hat{\sigma}_i = [(\hat{\sigma}_{i1}^2 + \hat{\sigma}_{i2}^2)/2]^{1/2}$. For studies that use dependent samples, an estimate of the variance of $\hat{\delta}_i$ is

$$\text{var}(\hat{\delta}_i) = [\hat{\delta}_i^2(\hat{\sigma}_{i1}^4+\hat{\sigma}_{i2}^4+2\hat{\rho}_i^2\hat{\sigma}_{i1}^2\hat{\sigma}_{i2}^2)/8\hat{\sigma}_i^4 df_i + (\hat{\sigma}_{i1}^2 + \hat{\sigma}_{i2}^2 - 2\hat{\rho}_i\hat{\sigma}_{i1}\hat{\sigma}_{i2})/\hat{\sigma}_i^2 df_i], \quad (6)$$

where $df_i = n_i - 1$. Equations 5 and 6 do not assume equal population variances within studies and follow from the results of Bonett (2008a). Although most studies that use dependent samples will report sample means and sample variances, the sample correlation needed in Equation 6 may not be reported. The sample correlation in study i is equal to $\hat{\rho}_i = [\hat{\sigma}_{i1}^2 + \hat{\sigma}_{i2}^2 - (df_i + 1)(\hat{\mu}_{i1} - \hat{\mu}_{i2})^2/t_i^2]/(2\hat{\sigma}_{i1}\hat{\sigma}_{i2})$, where t_i is the paired-samples t statistic in study i with df_i degrees of freedom.

The following approximate $100(1 - \alpha)\%$ Satterthwaite confidence interval for φ is proposed,

$$\bar{\varphi} \pm t_{\alpha/2;df}\left[m^{-2}\sum_{i=1}^{m}\text{var}(\hat{\varphi}_i)\right]^{1/2}, \quad (7)$$

where $\text{var}(\hat{\varphi}_i)$ is given by Equation 3 or 4 and $t_{\alpha/2;df}$ is a two-tailed critical value of a Student t distribution. For the case of independent samples within studies,

$$df = \left(\sum_{i=1}^{m}\sum_{j=1}^{2}\hat{\sigma}_{ij}^2/n_{ij}\right)^2 \Big/ \sum_{i=1}^{m}\sum_{j=1}^{2}\hat{\sigma}_{ij}^4/(n_{ij}^3 - n_{ij}^2), \quad (8)$$

and for the case of dependent samples within studies,

$$df = \left(\sum_{i=1}^{m}\hat{\sigma}_{di}^2/n_i\right)^2 \Big/ \sum_{i=1}^{m}\hat{\sigma}_{di}^4/(n_i^3 - n_i^2). \quad (9)$$

The following approximate $100(1 - \alpha)\%$ confidence interval for δ is proposed,

$$\bar{\delta} \pm z_{\alpha/2}\left[m^{-2}\sum_{i=1}^{m} b_i^2\,\text{var}(\hat{\delta}_i)\right]^{1/2}, \quad (10)$$

where $\text{var}(\hat{\delta}_i)$ is given by Equation 5 or 6 and $z_{\alpha/2}$ is a two-tailed critical z value. In the following section, the performance of Equations 7 and 10 is compared with the FE and RE methods of Bond et al. (2003) and Hedges and Vevea (1998).

Meta-analysis attempts to reproduce the results that would have been obtained if the raw data from all m studies

had been available. Equations 7 and 10 possess the important characteristic of exactly reproducing raw data results. In contrast, FE meta-analysis recommended by Bond et al. (2003) will reproduce raw data results from an $m \times 2$ analysis of variance only if the two-way interaction is assumed to be zero and the interaction sum of squares is pooled with the error term (Olkin & Sampson, 1998). Scheffé (1959, p. 126) described the use of no-interaction analysis of variance models as a "common but questionable practice" and explained why this practice should be avoided. Equation 10 exactly reproduces the raw data results that would be obtained with the confidence interval proposed by Bonett (2008a) in an $m \times 2$ design with m standardizers that are unique to each of the m levels. Note also that Equation 7 may be expressed as a standard Satterthwaite confidence interval for a linear contrast of unstandardized means (see, e.g., Maxwell & Delaney, 2004, pp. 300–301; Snedecor & Cochran, 1980, p. 228).

All the strengths and limitations of the Satterthwaite and Bonett (2008b) confidence intervals for linear contrast of means apply to Equations 7 and 10 in the context of a meta-analysis. Specifically, the Satterthwaite confidence interval performs well with highly unequal sample sizes and highly unequal population variances. The Satterthwaite confidence interval assumes that the response variable follows an approximate normal distribution, but its robustness to a violation of this assumption increases as the sample size per group increases. With samples sizes of 30 or more per group, the Satterthwaite confidence interval is remarkably robust to degrees of nonnormality that are typically encountered in practice (Bonett & Price, 2002).

Equation 10 requires stronger assumptions than Equation 7. Although Bonett (2008a) demonstrated that the confidence interval for standardized linear contrasts of means can tolerate population variances that are highly unequal, these standardized effects may not be meaningful measures of effect size unless the population variances are at most moderately unequal. With highly unequal population variances, an alternative standardizer recommended by Glass (1976) based on a variance estimate from a single control group or pretest condition may be more appropriate. If the Glass standardizer is used in the meta-analysis, Equations 5 and 6 are replaced with the variance estimates for the Glass standardizer given in Bonett (2008a). As explained in Bonett, Equation 10 does not share the robustness to nonnormality property of Equation 7, and increasing the sample size per group does not mitigate the problem. Furthermore, and perhaps more important, a meaningful interpretation of the population standardized effect size requires the response variable to be at most moderately nonnormal (Bonett, 2008a).

Equation 10 follows the tradition of averaging standardized effect sizes in which the effect size for study i has been standardized with sample variances from study i. Alternatively, one may obtain a confidence interval for $\varphi/[m^{-1}\sum_{i=1}^{m}\sum_{j=1}^{2}\sigma_{ij}^2/2]^{1/2}$, in which the average unstandardized effect is standardized with the variances from all m studies. This confidence interval is a special case of the confidence interval given by Bonett (2008a) and would be preferred to Equation 10 in applications in which the population variances, both between and within studies, are at most moderately unequal.

Given the known properties of Equations 7 and 10, Equation 10 should not be used unless the meta-analyst is convinced that the response variable is at most moderately nonnormal, and Equation 7 should not be used if the response variable is believed to be highly nonnormal and the sample sizes per group are very small. Evidence of approximate normality may be obtained from large-sample estimates of skewness and kurtosis reported in previous studies and not necessarily those studies used in the meta-analysis. In meta-analytic studies, evidence of nonnormality is often not provided by the researchers of the original work, and Equation 7 may be preferred to Equation 10 when sufficient evidence of approximate normality is lacking. Bond et al. (2003) provided additional justification for preferring a meta-analysis of unstandardized means over a meta-analysis of standardized means.

Modeling Effect-Size Heterogeneity

Bond et al. (2003) and Hedges and Vevea (1998) recommended a test of H_0: $\varphi_1 = \varphi_2 = \cdots = \varphi_m$ or H_0: $\delta_1 = \delta_2 = \cdots = \delta_m$ to assess effect-size heterogeneity. Statistical tests of these null hypotheses are routinely misused; specifically, failure to reject the null hypothesis does not imply homogeneity of effect size, and rejection of the null hypothesis does not imply that there are meaningfully large differences among the population effect sizes (see, e.g., Bonett & Wright, 2007). Furthermore, tests of effect-size homogeneity are incorrectly used to select between FE and RE models. It is common, although inappropriate, to select an FE method if the null hypothesis of homogeneity is not rejected and to select an RE method if the null hypothesis of homogeneity is rejected.

Although Equation 7 does not assume $\varphi_1 = \varphi_2 = \cdots = \varphi_m$ and Equation 10 does not assume $\delta_1 = \delta_2 = \cdots = \delta_m$, effect-size heterogeneity may be the result of differences in certain population characteristics that may moderate the effect sizes. For instance, differences in effect-size values may be due to differences in population demographics (e.g., age, education, social status) or specific aspects of the study (e.g., type of instructions, time limits, sex of experimenter). The m effect-size estimators may be expressed as a linear function of known population characteristics in the form of a general linear model,

$$\boldsymbol{Y} = \mathbf{X}\beta + \varepsilon, \tag{11}$$

where $\boldsymbol{Y}$ is an $m \times 1$ random vector with typical element $\hat{\varphi}_i$ or $b_i\hat{\delta}_i$, $\mathbf{X}$ is an $m \times q$ full-rank design matrix that codes

quantitative or qualitative differences among the m study populations, β is a $q \times 1$ vector of unknown parameters, and ε is an $m \times 1$ vector of random sampling errors with $\text{var}(\varepsilon_i)$ equal to $\text{var}(\hat{\varphi}_i)$ or $b_i^2\text{var}(\hat{\delta}_i)$ depending on the definition of $\boldsymbol{Y}$.

An ordinary least squares (OLS) estimator of β is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}, \tag{12}$$

with estimated covariance matrix

$$\text{cov}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{13}$$

where $\mathbf{V}$ is a diagonal matrix with $\text{var}(\varepsilon_i)$ in the ith diagonal element. The variance of $\hat{\beta}_k$ is the kth diagonal element of $\text{cov}(\hat{\beta})$, which will be denoted as $\text{var}(\hat{\beta}_k)$.

When modeling standardized mean differences, an approximate $100(1 - \alpha)\%$ confidence interval for β_k is

$$\hat{\beta}_k \pm z_{\alpha/2}[\text{var}(\hat{\beta}_k)]^{1/2}, \tag{14}$$

and when modeling unstandardized mean differences, an approximate $100(1 - \alpha)\%$ Satterthwaite confidence interval for β_k is

$$\hat{\beta}_k \pm t_{\alpha/2;df}[\text{var}(\hat{\beta}_k)]^{1/2}, \tag{15}$$

where α in Equations 14 and 15 may be replaced with α/ν to obtain ν simultaneous Bonferroni confidence intervals for any ν elements of β.

The value of df in Equation 15 depends on the type of design within each of the m studies. Let $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ with the kth row of $\mathbf{C}$ denoted as $\boldsymbol{c}$. Let c_i denote the ith element of $\boldsymbol{c}$. For dependent samples within studies

$$df = \left(\sum_{i=1}^{m} c_i\hat{\sigma}_{di}^2/n_i\right)^2 \Big/ \sum_{i=1}^{m} c_i^4\hat{\sigma}_{di}^4/(n_i^3 - n_i^2), \tag{16}$$

and for the case of independent samples within studies,

$$df = \left(\sum_{i=1}^{m}\sum_{j=1}^{2} c_{ij}^2\hat{\sigma}_{ij}^2/n_{ij}\right)^2 \Big/ \sum_{i=1}^{m}\sum_{j=1}^{2} c_{ij}^4\hat{\sigma}_{ij}^4/(n_{ij}^3 - n_{ij}^2), \tag{17}$$

where $c_{i1} = c_i$ and $c_{i2} = -c_i$.

The above approach differs from the estimated weighted least squares (EWLS) approach recommended by Hedges and Olkin (1985, p. 170), in which the weights are random variables. The EWLS estimator of β is $\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{Y}$, where $\mathbf{V}$ is defined in Equation 13 and $\text{cov}(\tilde{\beta}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. The EWLS estimator will be more efficient (i.e., have smaller variance) than the OLS estimator in large samples (Judge, Griffiths, Hill, Lütkepohl, & Lee, 1985, p. 170), and this fact has been the primary justification for the exclusive use of EWLS in FE meta-analysis. However, the performance of a confidence interval will depend on the mean square error of the estimator. The mean square error is defined as the sum of the squared estimator bias and the estimator variance. When the model (Equation 11) has not been perfectly specified (i.e., one or more necessary predictor variables have been omitted from the model), the EWLS estimator will be biased (see Appendix), and its mean square error can be considerably larger than the mean square error for the OLS estimator. Model misspecification is the rule rather than the exception, and the EWLS estimator may be difficult to justify in practice. If $\mathbf{X}$ in Equation 11 is set equal to an $m \times 1$ vector of ones, then β will equal δ or φ depending on the definition of $\boldsymbol{Y}$. Such a model assumes that the m population effect sizes are equal and effect-size heterogeneity represents a misspecification of the model. It follows that the classic FE estimators of δ or φ are EWLS estimators and will be more efficient but also more biased than the OLS estimators proposed here under effect-size heterogeneity when the expected values of the weights in the EWLS estimator are not all equal. See Bonett (2008b, Appendix B) for more details.

Note that $\hat{\beta}_k$ may be expressed as $\boldsymbol{c}'\boldsymbol{Y}$, a linear function of the m effect-size estimators. In some applications it will be more convenient to specify the elements of $\boldsymbol{c}$ directly to define an interesting linear contrast of effect sizes. For instance, with $m = 6$ studies in which the first two studies have been sampled from populations of college students and the other four studies have been sampled from noncollege populations, one might want to compute a confidence interval for the linear contrasts $(\varphi_1 + \varphi_2)/2 - (\varphi_3 + \varphi_4 + \varphi_5 + \varphi_6)/4$ or $(\delta_1 + \delta_2)/2 - (\delta_3 + \delta_4 + \delta_5 + \delta_6)/4$, which are defined by the contrast coefficients $c_1 = 1/2$, $c_2 = 1/2$, $c_3 = 1/4$, $c_4 = 1/4$, $c_5 = 1/4$, and $c_6 = 1/4$.

An approximate $100(1 - \alpha)\%$ confidence interval for $\sum_{i=1}^{m} c_i\delta_i$ is

$$\sum_{i=1}^{m} c_i\hat{\delta}_i \pm z_{\alpha/2}\left[\text{var}\left(\sum_{i=1}^{m} c_i\hat{\delta}_i\right)\right]^{1/2}, \tag{18}$$

where $\text{var}(\sum_{i=1}^{m} c_i\hat{\delta}_i) = \sum_{i=1}^{m} c_i^2\,\text{var}(\hat{\delta}_i)$. An approximate $100(1 - \alpha)\%$ Satterthwaite confidence interval for $\sum_{i=1}^{m} c_i\varphi_i$ is

$$\sum_{i=1}^{m} c_i\hat{\varphi}_i \pm t_{\alpha/2;df}\left[\text{var}\left(\sum_{i=1}^{m} c_i\hat{\varphi}_i\right)\right]^{1/2}, \tag{19}$$

where $\text{var}(\sum_{i=1}^{m} c_i\hat{\varphi}_i) = \sum_{i=1}^{m} c_i^2\,\text{var}(\hat{\varphi}_i)$ and α in Equations 18 and 19 may be replaced with α/ν to obtain ν simultaneous Bonferroni confidence intervals for any ν contrasts of interest. The value of df in Equation 19 is given by Equations 16 and 17 for designs with dependent samples and independent samples, respectively.

Monte Carlo Studies

The Monte Carlo method was used to compare the performance of Equations 7 and 10 with the competing methods of Bond et al. (2003) and Hedges and Vevea (1998). The Hedges–Vevea approach is general and may be applied to designs with independent or dependent samples. In contrast, the Bond–FE and Bond–RE methods were developed only for designs with independent samples. The Monte Carlo programs were written in GAUSS and executed on a Pentium IV computer. The Monte Carlo studies simulated the FE case in which the *m* studies cannot be assumed to be randomly selected from a superpopulation of studies.

Unstandardized Mean Differences

For the parameter $\varphi = m^{-1}\sum_{i=1}^{m}\varphi_i$, the performance of Equation 7 was compared with the Bond–FE and Bond–RE methods under 2,000 patterns of sample sizes and population effect sizes. This Monte Carlo study used homoscedastic normal samples within studies to accommodate the assumption of the Bond methods (recall that Equation 7 does not require homoscedasticity within studies). The coverage probabilities and confidence interval width were estimated from 50,000 Monte Carlo trials within each of the 2,000 conditions. To simplify the presentation of results, the coverage probability and confidence interval widths were averaged over four sets of 250 conditions for $m = 5$ and also for $m = 10$. The minimum coverage probability within each set of 250 conditions is also reported. The minimum coverage probability is perhaps the most important value because it shows how poorly a method can perform. In one of the four sets of conditions, the *m* sample sizes ranged from 20 to 40, and the *m* population effect sizes ranged from 0.25 to 0.75. In another set of 250 conditions, the sample sizes ranged from 20 to 40, but the population effect sizes were more disparate and ranged from 0.05 to 0.95. Other conditions used sample sizes that ranged from 40 to 80 across the *m* studies. In each of the 2,000 conditions, the population standard deviations ranged from 0.5 to 1.5 across the *m* studies (the Bond–FE and Bond–RE methods do not assume equal variances across studies).

The sample sizes, population effect sizes, and population variances for each of the 1,000 conditions were randomly generated from a uniform distribution within the specific range of values. For instance, in one of the conditions for $m = 5$, the computer-generated sample sizes might be [38 21 35 26 30], the computer-generated population effect sizes might be [0.68 0.31 0.54 0.58 0.43], and the computer-generated population standard deviations might be [1.2 0.7. 0.9 0.5 1.4]. After the computer-generated sample sizes and effect sizes were determined for each of the 2,000 conditions, a Monte Carlo study with 50,000 trials was conducted for each condition. For each condition, the coverage probability and confidence interval width were estimated for Equation 7 and Bond–FE and Bond–RE methods. The results for Equation 7 and the Bond methods for $1 - \alpha = .95$ are summarized in Table 1.

The best confidence interval method will have an average coverage probability close to .95 and a minimum coverage probability that is not too far below .95. If two methods have similar average and minimum coverage probabilities, then the method with the smallest average interval width is preferred. It can be seen from Table 1 that the Bond–FE method can have a coverage probability that is far below the nominal .95 level, rendering this method unacceptable for routine use. The Bond–RE method has better minimum coverage probabilities than the Bond–FE method; however, the average confidence interval width of the Bond–RE method is considerably larger than the average width of Equation 7.

Table 1
Performance Comparison of Three Meta-Analytic Confidence Intervals for Unstandardized Mean Differences: Independent Samples Within Studies

Sample size	Population effect size	Average coverage			Minimum coverage			Average width		
		Eq. 7	Bond–FE	Bond–RE	Eq. 7	Bond–FE	Bond–RE	Eq. 7	Bond–FE	Bond–RE
$m = 5$										
10–40	0.25–0.75	.950	.933	.960	.947	.664	.916	.559	.468	.740
	0.05–0.95	.950	.895	.973	.947	.285	.559	.559	.465	.906
20–80	0.25–0.75	.950	.915	.967	.947	.679	.900	.392	.321	.579
	0.05–0.95	.950	.836	.984	.947	.151	.913	.391	.322	.784
$m = 10$										
10–40	0.25–0.75	.950	.929	.963	.948	.753	.880	.392	.314	.425
	0.05–0.95	.950	.868	.975	.947	.183	.864	.391	.316	.530
20–80	0.25–0.75	.950	.904	.970	.947	.541	.886	.278	.225	.340
	0.05–0.95	.950	.824	.990	.947	.022	.918	.278	.226	.463

Note. Bond–FE is the fixed-effects method for unstandardized mean differences proposed by Bond et al. (2003), and Bond–RE is the random-effects method for unstandardized mean differences proposed by Bond et al. (2003). Eq. = Equation.

The results in Table 1 are consistent with the theoretical and simulation results reported by Bonett (2008b) for the case of Pearson correlations. The Bond–RE method is not expected to perform properly when the *m* studies are not a random sample from a large superpopulation of studies. The results in Table 1 illustrate the poor performance of the Bond–RE method when this method is inappropriately used to accommodate unequal population effect sizes in nonrandomly selected studies. If the sample sizes and population variances are unequal within studies, Equation 7 continues to perform properly, but the performance of the Bond method will be worse than what is shown in Table 1. If the sample sizes or population variances are more heterogeneous across studies than those examined here, then the performance of the Bond methods will be worse than what is shown in Table 1. Equation 7 continues to perform well when the population variances and sample sizes within studies are highly unequal, whereas the Bond–FE method breaks down further under these conditions. Although not shown in Table 1, it was found that the performance of Equation 7 with dependent samples within studies exhibited excellent performance characteristics that were nearly identical to those reported for Equation 7 in Table 1. Recall that the Bond–FE and Bond–RE methods are appropriate only for studies with independent samples.

Standardized Mean Differences

For the parameter $\delta = m^{-1}\sum_{i=1}^{m}\delta_i$, the performance of Equation 10 was compared with the FE and RE methods proposed by Hedges and Vevea (1998) with either independent or dependent samples within each study. Although Hedges and Vevea illustrated their methods using the standardized mean difference and its variance proposed by Hedges (1981), which assumes equal population variances within each study, the Hedges–Vevea approach is general and may be used with the standardized mean difference and its variance recommended by Bonett (2008a), which does not assume equal population variances with each study. The standardized mean difference and its variance recommended by Bonett were used in Equation 10 and the Hedges–Vevea methods. Four thousand patterns of sample sizes and population effect sizes were examined, with 2,000 patterns for the case of independent samples and 2,000 patterns for the case of dependent samples. The performance of Equation 10 and the Hedges–Vevea methods are invariant with respect to heteroscedasticity across studies, and it was not necessary to vary the population variances across the *m* studies. However, heterogeneity of population correlations across the *m* studies in the case of dependent samples will affect the performance of the Hedges–Vevea confidence intervals. In each of the 2,000 conditions for a given value of *m*, the population correlations were selected from a range of 0.5–0.8 across the *m* studies. Unlike the variance of an unstandardized difference, the variance of the standardized difference depends on the magnitude of the population effect size within each study. For this reason, moderately disparate and highly disparate ranges of population effect sizes were examined for both small (−0.2 to 0.2 and −0.4 to 0.4) and large (0.8 to 1.2 and 0.6 to 1.4) population effect sizes. The Monte Carlo results for the case of independent samples within studies are summarized in Table 2, and the Monte Carlo results for the case of dependent samples within studies are summarized in Table 3.

It can be seen from Tables 2 and 3 that the Hedges–Vevea methods can have a coverage probability that is far below the nominal .95 level. The average confidence interval width was considerably larger for the Hedges–Vevea RE methods than for Equation 10. The results in Tables 2 and 3 are also consistent with theoretical and simulation results reported by Bonett (2008b) for the case of Pearson correlations. If the sample sizes (or population correlations for the case of dependent samples) are more heterogeneous across studies than those examined here, then the performance of the Hedges–Vevea methods will be worse than what is shown in Table 1, whereas the excellent performance of Equation 10 remains essentially unchanged.

Equations 14, 15, 18, and 19 are all special cases of the general Satterthwaite or Bonett (2008a) confidence intervals, which are known to have excellent performance characteristics. The EWLS competitors to Equations 14 and 15 will exhibit poor performance, similar to the FE methods in Tables 1–3, under realistic conditions in which the sample sizes are unequal across studies and the linear model has not been perfectly specified.

An Alternative View of Meta-Analysis

Meta-analysis has been used primarily as a way to summarize the results of a large number of published studies. These meta-analytic reviews typically attempt to assimilate the results of as many studies as possible, despite the fact that the studies may differ dramatically in the quality of their sampling designs, research designs, or psychometric properties of key variables. Eysenck (1978) described this use of meta-analysis as "an exercise in megasilliness." The view taken here is that meta-analysis is most appropriately applied to a small number of carefully selected studies of the highest quality with the primary goals of obtaining accurate estimates of effect size and detecting important moderator effects.

Meta-analysis should no longer be viewed only as a method of secondary data analysis. The confidence intervals proposed here (Equations 7, 10, 14, 15, 18, and 19) and in Bonett (2008b) may also be applied in new studies that integrate the results of previous studies. For instance, suppose a recently published study reported that a new treatment for anxiety was found to be superior to a standard

Table 2
Performance Comparison of Three Meta-Analytic Confidence Intervals for Standardized Mean Differences: Independent Samples Within Studies

Sample size	Population effect size	Average coverage			Minimum coverage			Average width		
		Eq. 10	HV–FE	HV–RE	Eq. 10	HV–FE	HV–RE	Eq. 10	HV–FE	HV–RE
$m = 5$										
10–40	−0.20–0.20	.952	.953	.969	.949	.940	.960	.549	.510	.588
	−0.40–0.40	.951	.951	.975	.949	.911	.953	.541	.508	.652
	0.80–1.20	.951	.951	.966	.948	.923	.953	.577	.540	.620
	0.60–1.40	.951	.943	.971	.948	.858	.942	.565	.545	.689
20–80	−0.20–0.20	.951	.949	.969	.948	.922	.955	.381	.359	.604
	−0.40–0.40	.951	.940	.979	.948	.823	.938	.386	.363	.529
	0.80–1.20	.951	.947	.967	.948	.898	.942	.405	.381	.454
	0.60–1.40	.951	.937	.978	.948	.752	.940	.406	.382	.546
$m = 10$										
10–40	−0.20–0.20	.952	.954	.968	.948	.927	.955	.387	.361	.405
	−0.40–0.40	.952	.948	.978	.949	.838	.955	.384	.357	.455
	0.80–1.20	.951	.947	.963	.949	.922	.949	.406	.378	.422
	0.60–1.40	.951	.939	.972	.948	.838	.937	.409	.380	.472
20–80	−0.20–0.20	.951	.949	.970	.948	.902	.944	.270	.251	.296
	−0.40–0.40	.951	.939	.987	.949	.849	.953	.271	.253	.375
	0.80–1.20	.951	.945	.967	.949	.886	.948	.286	.266	.310
	0.60–1.40	.951	.929	.983	.948	.784	.935	.287	.267	.386

Note. HV–FE is the fixed-effects method for standardized mean differences proposed by Hedges and Vevea (1998), and HV–RE is the random-effects method for standardized mean differences proposed by Hedges and Vevea (1998). Eq. = Equation.

treatment in a population of college women. A new study is conducted by another researcher with the goal of determining whether this new treatment is also effective in a population of college men. The researcher of the new study might attempt to replicate and extend the results of the published study using a stratified random sample of both female and male participants. The current practice is simply to cite the results of the published study and then to report the results of the new study. An alternative approach would be to combine the results of the published study with the results of the new study. Before combining the results (using Equation 7 or 10), the researcher would first assess the degree of effect-size heterogeneity (using Equation 18 or 19) to determine whether the effect of treatment for women has been replicated and to assess the magnitude of the moderating effect of sex. If the effect of treatment for women can be replicated in the new study and if the moderating effect of sex is not too large, then Equation 7 or 10 could be used to describe the average effect of treatment across the three study populations (i.e., the two female populations and the one male population). Future studies that investigate the effect of this new treatment for anxiety in other types of populations or with modifications to the new treatment would incorporate the results from all relevant previous studies. With this approach, increasingly accurate estimates of the treatment effect size will be obtained as well as important information regarding possible moderating effects of the treatment. With this alternative view of meta-analysis, the fundamental process of the scientific method will be more closely approximated by researchers who incorporate prior empirical information into their studies, "add a modicum of new and better data to it, and thereby advance toward an ever more profound, complete, and accurate explanation of reality" (Hunt, 1997, p. 1).

Traditional meta-analysis has focused on combining standardized or unstandardized mean differences rather than individual means. If meta-analysis is used to incorporate the results of prior research into a new study, it will often be necessary to combine individual means for a common treatment condition from prior studies and combine the average prior mean with an appropriate mean in the new study. For instance, the means in the control condition from three previous studies could be combined with the control group mean of a new study that compares a treatment with a control condition. The researcher could then compute confidence intervals for $(\mu_1 + \mu_2 + \mu_3)/3 - \mu_4$ and $(\mu_1 + \mu_2 + \mu_3 + \mu_4)/4 - \mu_5$, where μ_1, μ_2, and μ_3 are the population means under a control condition estimated from the three previous studies, μ_4 is the population mean under a control condition estimated from the new study, and μ_5 is the population mean under the treatment condition estimated from the new study. The first confidence interval provides evidence of control group replication, and the second confidence interval provides evidence of a treatment effect. Confidence intervals for these linear contrasts of population means are computed with the standard Satterth-

Table 3
Performance Comparison of Three Meta-Analytic Confidence Intervals for Standardized Mean Differences: Dependent Samples Within Studies

Sample size	Population effect size	Average coverage			Minimum coverage			Average width		
		Eq. 10	HV–FE	HV–RE	Eq. 10	HV–FE	HV–RE	Eq. 10	HV–FE	HV–RE
$m = 5$										
10–40	−0.20–0.20	.954	.925	.966	.951	.867	.939	.334	.289	.366
	−0.40–0.40	.954	.917	.981	.950	.649	.916	.333	.290	.473
	0.80–1.20	.949	.901	.951	.945	.757	.903	.400	.359	.435
	0.60–1.40	.949	.843	.967	.944	.348	.883	.408	.362	.537
20–80	−0.20–0.20	.952	.931	.973	.949	.746	.919	.229	.204	.285
	−0.40–0.40	.952	.885	.991	.949	.328	.908	.230	.206	.435
	0.80–1.20	.950	.913	.965	.947	.665	.910	.280	.256	.334
	0.60–1.40	.949	.834	.981	.946	.236	.904	.283	.256	.446
$m = 10$										
10–40	−0.20–0.20	.954	.938	.970	.951	.866	.932	.235	.200	.249
	−0.40–0.40	.954	.913	.990	.951	.676	.940	.238	.203	.342
	0.80–1.20	.949	.860	.937	.946	.712	.869	.287	.254	.301
	0.60–1.40	.949	.754	.967	.944	.309	.833	.288	.252	.377
20–80	−0.20–0.20	.952	.926	.981	.950	.678	.931	.161	.142	.202
	−0.40–0.40	.952	.874	.998	.949	.302	.960	.162	.143	.313
	0.80–1.20	.950	.887	.963	.946	.664	.903	.199	.180	.232
	0.60–1.40	.950	.742	.990	.946	.196	.890	.202	.181	.334

Note. HV–FE is the fixed-effects method for standardized mean differences proposed by Hedges and Vevea (1998), and HV–RE is the random-effects method for standardized mean differences proposed by Hedges and Vevea (1998). Eq. = Equation.

waite confidence interval (see, e.g., Maxwell & Delaney, 2004, pp. 300–301; Snedecor & Cochran, 1980, p. 228).

Examples

Three examples, with hypothetical data, illustrate the application of the meta-analytic confidence intervals proposed here. The first example is a meta-analysis of four two-group experiments. The second example is a meta-analysis of five pretest–posttest studies. The third example illustrates the use of meta-analysis to integrate the results of previous studies into a new study.

Example 1

Four eyewitness identification studies assessed participants' certainty in their selection of a target individual from a photo lineup. In all four studies, participants viewed a video in which the target individual could be seen. The four studies used a variety of treatment conditions but had two treatment conditions in common. In one common condition, the participants were told that the target individual "will be" in the photo lineup, and in the second common condition, participants were told that the target individual "might be" in the photo lineup. Two of the four studies used $x = 5$ photos in the lineup, and the other two studies used $x = 7$ photos in the lineup. All four studies sampled from study populations of volunteer introductory psychology students. The sample means and standard deviations from the four studies are presented in Table 4. These sample means and standard deviations were used to compute the standardized effect sizes and their variances, which are also reported in Table 4.

A confidence interval for a linear contrast of population standardized effect sizes $(\delta_1 + \delta_2)/2 - (\delta_3 + \delta_4)/2$ will

Table 4
Summary Information for Example 1

Study	n_1	n_2	x	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\delta}_1$	$\text{var}(\hat{\delta}_1)$
1	40	40	5	7.4	6.3	1.7	2.3	0.539	0.052
2	20	20	5	6.9	5.7	1.5	2.0	0.665	0.107
3	25	25	7	6.8	5.8	1.6	1.8	0.578	0.084
4	30	30	7	6.6	5.5	1.8	2.1	0.555	0.070

Note. n_1, $\hat{\mu}_1$, and $\hat{\sigma}_1$ are the sample size, sample mean, and sample standard deviation in the "will be" instruction condition; n_2, $\hat{\mu}_2$, and $\hat{\sigma}_2$ are the sample size, sample mean, and sample standard deviation in the "might be" instruction condition; x is the number of photos in the lineup.

provide information about the magnitude of the moderating effect of the number of photos in the lineup. The point estimate of this contrast is 0.036, with a standard error of 0.279. The 95% confidence interval for the contrast is −0.51 to 0.58. This confidence includes zero but is too wide to assess the magnitude of the moderating effect accurately. Additional research is needed to obtain a more accurate assessment of the moderating effect of the number of photos. The point estimate of $\delta = (\delta_1 + \delta_2 + \delta_3 + \delta_4)/4$ is 0.584, with a standard error of 0.140, indicating greater certainty in the will-be than the might-be condition. The 95% confidence interval for δ is 0.31 to 0.86, which could be interpreted as a small to moderate effect of the type of instruction on an eyewitness's certainty of selecting the target person from a photo lineup. This result applies to the four study populations of introductory psychology students.

Example 2

Five published studies employed a pretest–posttest design and reported the effect of relaxation therapy on hours of migraine headaches per week. The response variable metric (hours per week) is well understood, and unstandardized effects will be analyzed. The number of weeks of relaxation therapy (x) varied across studies and is a potentially interesting moderator variable. All five studies sampled from study populations of adults who responded to a newspaper ad request for volunteers. The necessary sample statistics (means, standard deviations, correlations) were extracted from the five studies. These sample statistics were used to compute the unstandardized effect sizes and their variances for each study. The sample statistics and effect size information are summarized in Table 5.

To assess the magnitude of the moderating effect of relaxation therapy duration on the unstandardized effect size, set the design matrix $\mathbf{X}$ in Equation 11 to a 5×2 matrix with ones in the first column and the values 2, 3, 3, 4, and 4 in the second column. The estimated population slope is 0.85, with a standard error of 0.50 and degree of freedom of 87.07. The 95% Satterthwaite confidence interval for the population slope is −0.144 to 1.84. This interval includes zero but may be too wide to conclude that the moderating effects of therapy duration is absent, because if the slope was as large as 1.84, some experts would consider this to be a nontrivial clinical effect. Additional studies must be included in another meta-analysis to obtain a more precise estimate of the population slope.

Given the inconclusive result regarding the moderating effect of therapy duration, the average effect size in the five study populations will be of interest. An estimate of $\varphi = (\varphi_1 + \varphi_2 + \cdots + \varphi_5)/5$ is 10.7 hr per week, with a standard error of 0.44 and degree of freedom of 83.08. The 95% confidence interval for φ is 9.85 to 11.60 and suggests that 2–4 weeks of relaxation therapy would decrease the mean hours of migraine headaches per week by 9.85 to 11.60 in the five study populations of adult volunteers.

Example 3

A researcher conducted a two-group experiment that compared the effect of 100 mg of a new antidepressant drug with a placebo in a study population of patients who had been diagnosed with mild to moderate levels of depression. The researcher used the Beck Depression Inventory (BDI-II) to measure depression after treatment. The scale of the BDI-II is well understood among clinicians, and the researcher decided to analyze unstandardized effect sizes. The researcher used the placebo group means and standard deviations from three previously published that also sampled from populations of patients with mild to moderate levels of depression. These three previous studies compared a placebo condition with other treatment conditions that were not relevant to the current study. The researcher obtained a random sample of 80 patients with mild to moderate depression and randomly assigned 20 patients to a placebo condition and 60 patients to the 100-mg drug condition. The results from the three previous studies and the new study (Study 4) are given in the first four rows of Table 6.

The contrast $(\mu_1 + \mu_2 + \mu_3)/3 - \mu_4$ provides a placebo replication check, and the contrast $(\mu_1 + \mu_2 + \mu_3 + \mu_4)/4 - \mu_5$ assesses the effect of the new drug, where μ_5 is the population mean BDI-II score under a 100-mg drug condition. Application of the standard 95% Satterthwaite confidence interval for a linear contrast of means with a Bonferroni adjustment (see, e.g., Maxwell & Delaney, 2004, pp. 300–301; Snedecor & Cochran, 1980, p. 228)

Table 5
Summary Information for Example 2

Study	n	x	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\rho}$	$\hat{\varphi}_1$	var($\hat{\varphi}_1$)
1	45	2	20.1	10.4	9.3	7.8	.87	9.7	0.469
2	15	3	20.5	10.2	9.9	8.0	.92	10.3	1.085
3	20	3	19.3	8.5	10.1	8.4	.85	10.8	1.417
4	20	4	21.5	10.3	10.5	8.1	.90	11.2	1.139
5	30	4	19.4	7.8	9.8	8.7	.88	11.6	0.722

Note. n is the sample size, x is the number of weeks of relaxation therapy, and $\hat{\rho}$ is the sample Pearson correlation; $\hat{\mu}_1$ and $\hat{\sigma}_1$ are the pretest mean and standard deviation; $\hat{\mu}_2$ and $\hat{\sigma}_2$ are the posttest mean and standard deviation.

Table 6
Summary Information for Example 3

Study	Placebo			50 mg			100 mg		
	n	$\hat{\mu}$	$\hat{\sigma}$	n	$\hat{\mu}$	$\hat{\sigma}$	n	$\hat{\mu}$	$\hat{\sigma}$
1	60	21.1	5.1	—	—	—	—	—	—
2	80	18.9	4.7	—	—	—	—	—	—
3	50	22.4	5.6	—	—	—	—	—	—
4	20	20.9	4.8	—	—	—	60	11.2	4.1
5	30	22.0	5.9	40	15.3	4.0	40	10.7	4.2

Note. n, $\hat{\mu}$, and $\hat{\sigma}$ are the sample size, sample mean, and sample standard deviation. Dashes indicate data not obtained.

gives −2.84 to 2.64 for the first contrast and 8.13 to 11.12 for the second contrast. The confidence interval for the first contrast includes zero and is sufficiently narrow to support a claim of replication in the placebo condition.[1] The confidence interval for the second contrast provides evidence of treatment effectiveness, and the researcher can state with 95% confidence that the mean BDI-II score in the study population of depressed patients would be 8.13 to 11.12 lower if they had all received 100 mg of the new drug instead of a placebo. If the researcher had followed current practice and analyzed the data only from Study 4, the 95% Satterthwaite confidence interval for $\mu_4 - \mu_5$ would be 7.25 to 12.15, which is wider than the meta-analytic result, and the study would not have provided any evidence of placebo replication.

Suppose the meta-analytic results of Studies 1–4 are published and another researcher wanted to compare the effectiveness of the new drug under both 50-mg and 100-mg drug conditions. The second researcher obtained a random sample of 110 patients with mild to moderate depression and randomly assigned 30 patients to a placebo condition, 40 patients to a 50-mg condition, and 40 patients to a 100-mg condition. The results from the four previous studies and the new study (Study 5) are given in Table 6.

Confidence intervals for $(\mu_1 + \mu_2 + \mu_3 + \mu_4)/4 - \mu_6$ and $\mu_5 - \mu_8$ provide evidence of placebo and 100-mg replication, respectively, where μ_1, μ_2, μ_3, μ_4, and μ_6 are the population mean BDI-II scores under the placebo conditions of Studies 1–5 and μ_5 and μ_8 are the population mean BDI-II scores under the 100-mg conditions of Studies 4 and 5. Confidence intervals for $(\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_6)/5 - \mu_7$ and $(\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_6)/5 - (\mu_5 + \mu_8)/2$ compare a 50-mg treatment with a placebo and a 100-mg treatment with a placebo, respectively. Simultaneous ($v = 4$) 95% Satterthwaite confidence intervals are −4.19 to 1.84 for the placebo replication and −1.67 to 2.66 for the 100-mg replication. Both intervals include zero and are sufficiently narrow to support a claim of replication in both the placebo and 100-mg treatment conditions. The other two simultaneous 95% Satterthwaite confidence intervals are 3.86 to 7.66 for the 50-mg versus placebo comparison and 8.67 to 11.55 for the 100-mg versus placebo comparison. The researcher can state with 95% confidence that the mean BDI-II score in the study populations of depressed patients would be 3.86 to 7.66 lower if they had all received 50 mg of the new drug instead of a placebo and 8.67 to 11.55 lower if they had all received 100 mg of the new drug instead of a placebo.

Future studies could extend this line of research in several ways. In each study, the researcher would attempt to replicate previous results and also extend or clarify the theory by including additional levels of a particular factor or introducing new factors into the study. For instance, one future study could examine four drug dosages, and another future study could extend the results further with a 2×4 factorial experiment using the same four drug dosages and a second two-level factor that compares a drug-only treatment with a drug plus cognitive behavioral therapy treatment. In each study, the researcher attempts to replicate previous findings and extend the theory in important new directions.

Conclusion

Equation 7 is robust to moderate nonnormality, and its robustness to more extreme degrees of nonnormality increases with larger sample sizes per study. Equation 10 is not robust to nonnormality. When analyzing raw data, the researcher can employ a wide variety of data transformations to reduce nonnormality, but the meta-analyst may not have access to the raw data and must rely on diagnostic information reported within each study to assess the plausibility of the normality assumption.

If the response variable has a well-understood metric but has a highly skewed distribution, a meta-analysis of medians would be preferred to Equation 7 or 10. The average median difference is $\gamma = m^{-1}\sum_{i=1}^{m}(\gamma_{i1} - \gamma_{i2})$, where γ_{ij} is the population median under treatment j in study i. An approximate $100(1 - \alpha)\%$ confidence interval for γ is

$$\hat{\gamma} \pm z_{\alpha/2}\left\{m^{-2}\sum_{i=1}^{m}[\mathrm{var}(\hat{\gamma}_{i1}) + \mathrm{var}(\hat{\gamma}_{i2})]\right\}^{1/2}, \qquad (20)$$

where $\mathrm{var}(\hat{\gamma}_{ij})$ is given by Bonett and Price (2002) and $\hat{\gamma} = m^{-1}\sum_{i=1}^{m}(\hat{\gamma}_{i1} - \hat{\gamma}_{i2})$. However, $\mathrm{var}(\hat{\gamma}_{ij})$ is computed

[1] Some researchers may want to make a distinction between *weak replication* evidence in which the confidence interval includes zero but the confidence interval is wide and *strong replication* evidence in which the confidence interval suggests that the replication contrast is small. Strong replication does not require the confidence interval to include zero, but the lower or upper limit, whichever is further from zero, must be close enough to zero to suggest that the magnitude of population replication contrast is arguably trivial or unimportant.

from the raw data rather than summary statistics, and therefore it is incumbent on the researchers of the original studies to report $\hat{\gamma}_{ij}$ and var($\hat{\gamma}_{ij}$) when the response variable is highly skewed. Equation 20 is a special case of the confidence interval for a general linear function of medians given by Bonett and Price.

If the response variable has a well-understood metric but has an approximately symmetric and highly leptokurtic (heavy-tailed) distribution, a meta-analysis of trimmed means would be preferred to Equation 7 or 10. Wilcox (2005, p. 290) described a confidence interval method for a linear contrast of trimmed means. Trimmed means and their variances must be computed from the raw data and cannot be deduced from the sample means and variances that are typically reported. Thus, a meta-analysis of trimmed means is limited to a synthesis of *m* studies that have reported trimmed means and their variances. When the response variable distribution is symmetric, both the sample mean and the sample trimmed mean estimate the mean of the study population. A trimmed mean should be used with caution when the response variable is skewed because the sample trimmed mean then estimates the mean of a subset of the study population, which leads to problems of interpretation and a reduction in external validity.

The analytical expression for the bias of a weighted average of correlations given by Bonett (2008b) may be extended in an obvious way to the case of standardized or unstandardized mean differences considered here. The analytical results for estimator bias and the simulation results presented here and in Bonett (2008b) provide additional evidence of the unacceptable performance of the classic FE methods under realistic conditions of effect-size and sample-size heterogeneity. These findings further support the recommendations of the National Research Council (1992) and Hunter and Schmidt (2000) to discontinue the use of the classic FE methods.

The RE methods have been recommended as alternatives to the FE methods because RE effects methods do not make the unrealistic assumption of effect-size homogeneity (Hunter & Schmidt, 2000), and current research in meta-analysis methods appears to be focused on developing new RE methods. However, the RE methods require an unrealistic assumption that the *m* studies have been randomly sampled from a large and clearly defined superpopulation of studies. The random sample assumption will be virtually impossible to satisfy in the typical meta-analysis in which studies are published sequentially over time with more recent articles intentionally designed to be similar or dissimilar to previous studies. Even if the random sample assumption could be justified in certain applications, the standard deviation of the distribution of population effect sizes becomes a key parameter to estimate, and a very large number of studies will be required to obtain a sufficiently narrow confidence interval for the effect-size standard deviation in the superpopulation. Given the implausibility that the *m* studies are a true random sample from a well-defined superpopulation and the fundamental limitations of interval estimation of the random effect-size standard deviation, RE meta-analysis methods cannot be recommended for routine use.

The new FE methods presented here and in Bonett (2008b) do not assume effect-size homogeneity; nor do they assume that the *m* studies have been randomly sampled from a superpopulation of studies. The methods for assessing effect-size heterogeneity presented here and in Bonett have excellent performance characteristics and provide important information regarding the nature and magnitude of effect-size heterogeneity. Shapiro (1994, p. 771) lamented that current meta-analysis methods involve "computer models of bewildering complexity." The meta-analysis methods presented here and in Bonett are both conceptually and computationally simple.

An alternative view of meta-analysis is proposed here, consistent with the recommendation of Slavin (1986), in which meta-analysis is applied to a small number of carefully selected and high-quality studies. The alternative view also extends the use of meta-analysis as a general methodology for incorporating the results of previous research into a new study. Adoption of this new methodology will require researchers to radically change the way they design and statistically analyze their studies. Incorporating the results of previous studies into each new study provides a formal mechanism for simultaneously replicating and extending research findings and has the potential of reducing the "chaos" (Hunt, 1997, pp. 1–19) that exemplifies current research in behavioral, social, and medical research.

References

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8,* 406–418.

Bonett, D. G. (2008a). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13,* 99–109.

Bonett, D. G. (2008b). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods, 13,* 173–181.

Bonett, D. G., & Price, R. M. (2002). Statistical inference for linear function of medians: Confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods, 7,* 380–383.

Bonett, D. G., & Wright, T. A. (2007). Comments and recommendations regarding the hypothesis testing controversy. *Journal of Organizational Behavior, 28,* 647–659.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33,* 517.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Goldberger, A. S. (1991). *A course in econometrics.* Cambridge, MA: Harvard University Press.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 7,* 107–128.

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

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

Hunt, M. (1997). *How science takes stock: The story of meta-analysis.* New York: Russell Sage Foundation.

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

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., & Lee, T.-C. (1985). *The theory and practice of econometrics* (2nd ed.). New York: Wiley.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

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

Olkin, I., & Sampson, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics, 54,* 317–322.

Raudenbush, S. W. (1994). Random effects models. In H. Copper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches.* Göttingen, Germany: Hogrefe & Huber.

Shapiro, S. (1994). Meta-analysis/shmeta-analysis. *American Journal of Epidemiology, 140,* 771–778.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15,* 5–11.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames, IA: Iowa State University Press.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26,* 37–52.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). Burlington, MA: Elsevier.

(*Appendix follows*)

Appendix

Bias of Weighted Least Squares (WLS) Estimators

Consider the linear model

$$\boldsymbol{Y} = \mathbf{X}\beta + \mathbf{W}\theta + \varepsilon, \tag{A1}$$

where the columns of $\mathbf{X}$ and $\mathbf{W}$ represent fixed predictors of $\boldsymbol{Y}$. Assume that $\text{cov}(\varepsilon) = \mathbf{V}$ and $E(\varepsilon) = 0$. The assumption $E(\varepsilon) = 0$ implies $E(\mathbf{Y}) = \mathbf{X}\beta + \mathbf{W}\theta$. For simplicity of presentation, it will be assumed that $\mathbf{V}$ is a diagonal matrix of known constants. In meta-analytic applications, $\mathbf{V}$ is often a matrix of random variables that are estimated from sample data. The results presented here for the WLS estimator ($\mathbf{V}$ known) will apply to estimated weighted least squares (EWLS) estimators ($\mathbf{V}$ random) in large samples. Assume $\mathbf{X}'\mathbf{W} = 0$, that is, the columns of $\mathbf{W}$ are orthogonal to the columns of $\mathbf{X}$. This assumption can be satisfied without altering the value of θ by replacing $\mathbf{W}$ in Equation A1 with the orthogonalized predictors $\mathbf{W} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$.

Suppose that Equation A1 is the correct model and β is estimated from the following misspecified model:

$$\boldsymbol{Y} = \mathbf{X}\beta + \varepsilon. \tag{A2}$$

The WLS estimator of β is $\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{Y}$, and its bias is equal to

$$\begin{aligned} E(\tilde{\beta} - \beta) &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{E}(\boldsymbol{Y}) - \beta \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{X}\beta + \mathbf{W}\theta) - \beta \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{W}\theta. \end{aligned}$$

The bias will be zero if $\theta = 0$ (i.e., the model has actually not been misspecified) or if $\mathbf{X}'\mathbf{V}^{-1}\mathbf{W} = 0$. Given $\mathbf{X}'\mathbf{W} = 0$, $\mathbf{X}'\mathbf{V}^{-1}\mathbf{W}$ will not generally equal a null matrix unless the elements in $\mathbf{V}$ are equal. The elements of $\mathbf{V}$ are a function of the sample sizes, which are typically unequal in meta-analytic studies. The bias of the WLS estimator does not vanish as the sample size is increased. Additional small-sample bias may be introduced in EWLS estimators when $\mathbf{V}$ is random. Unlike the WLS estimator, the OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y}$ is unbiased:

$$\begin{aligned} E(\hat{\beta} - \beta) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}(\boldsymbol{Y}) - \beta \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{W}\theta) - \beta \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\theta \\ &= 0. \end{aligned}$$

It can be shown that when $\mathbf{X}$ in Equation A2 is an $m \times 1$ vector of ones and $\boldsymbol{Y}$ contains effect-size estimators, the meta-analytic FE estimators are EWLS estimators, and effect-size heterogeneity is a model misspecification. It is the bias of the classic FE estimators that is responsible for the poor performance of the FE confidence intervals in Tables 1–3.

Received June 18, 2008
Revision received January 26, 2009
Accepted March 24, 2009 ■