Psychological Methods
1998, Vol. 3, No. 4, 486–504

1082-989X/98/$3.00

Fixed- and Random-Effects Models in Meta-Analysis

Larry V. Hedges
University of Chicago

Jack L. Vevea
University of North Carolina at Chapel Hill

There are 2 families of statistical procedures in meta-analysis: fixed- and random-effects procedures. They were developed for somewhat different inference goals: making inferences about the effect parameters in the studies that have been observed versus making inferences about the distribution of effect parameters in a population of studies from a random sample of studies. The authors evaluate the performance of confidence intervals and hypothesis tests when each type of statistical procedure is used for each type of inference and confirm that each procedure is best for making the kind of inference for which it was designed. Conditionally random-effects procedures (a hybrid type) are shown to have properties in between those of fixed- and random-effects procedures.

The use of quantitative methods to summarize the results of several empirical research studies, or *meta-analysis,* is now widely used in psychology, medicine, and the social sciences. Meta-analysis usually involves describing the results of each study by means of a numerical index (an estimate of effect size, such as a correlation coefficient, a standardized mean difference, or an odds ratio) and then combining these estimates across studies to obtain a summary. Two somewhat different statistical models have been developed for inference about average effect size from a collection of studies, called the *fixed-effects* and *random-effects* models. (A third alternative, the *mixed-effects* model, arises in conjunction with analyses involving study-level covariates or moderator variables, which we do not consider in this article; see Hedges, 1992.)

Fixed-effects models treat the effect-size parameters as fixed but unknown constants to be estimated and usually (but not necessarily) are used in conjunction with assumptions about the homogeneity of effect parameters (see, e.g., Hedges, 1982; Rosenthal & Rubin, 1982). Random-effects models treat the effect-size parameters as if they were a random sample from a population of effect parameters and estimate hyperparameters (usually just the mean and variance) describing this population of effect parameters (see, e.g., DerSimonian & Laird, 1986; Hedges, 1983; Schmidt & Hunter, 1977). Although inference procedures based on these models have been available for well over a decade, there is still considerable confusion about the differences between them.

The purpose of this article is to clarify the conceptual distinctions between the models themselves and provide analytic information about the properties of statistical procedures (significance tests and confidence intervals) based on them. First, we discuss fixed- and random-effects models for meta-analysis in conjunction with fixed- and random-effects models in the more familiar context of analysis of variance (ANOVA), emphasizing that choice of model depends on the inferences the analyst wishes to make. Next, we develop the sampling properties of fixed- and random-effects procedures in detail, followed by those of a hybrid, which we call *conditionally random-effects* procedures. Then we compare the rejection rates of fixed-, random-, and conditionally random-effects tests analytically and address the adequacy of confidence intervals developed by means of these three procedures. Finally, we offer suggestions about applications of these methods in meta-analysis.

Fixed-Effects Versus Random-Effects Models in Meta-Analysis

There has been a great deal of confusion about the difference between fixed- and random-effects models

Larry V. Hedges, Departments of Education, Psychology, and Sociology, University of Chicago; Jack L. Vevea, Department of Psychology, University of North Carolina at Chapel Hill.

Correspondence concerning this article should be addressed to Larry V. Hedges, Department of Education, University of Chicago, Chicago, Illinois 60637. Electronic mail may be sent to hedge@cicero.spc.uchicago.edu.

in meta-analysis. Because this is part of a larger issue of conditional versus unconditional analyses in statistics, differences of opinion about the appropriateness of these analyses are likely to persist for some time (see Camilli, 1990, for a discussion of the conditionality issue in another context). However, we attempt here to clarify some of the issues.

The choice between fixed- and random-effects procedures has sometimes been framed as entirely a question of homogeneity of the effect-size parameters. That is, if all of the studies estimate a common effect-size parameter, then fixed-effects analyses are appropriate. However, if there is evidence of heterogeneity among the population effects estimated by the various studies, then random-effects procedures should be used. Although it is true that fixed- and random-effects analyses give similar answers when there is in fact a common population effect size, the inference models remain distinct. Moreover, there may be situations in which the fixed-effects analysis is appropriate even when there is substantial heterogeneity of results (e.g., when the question is specifically about a particular set of studies that have already been conducted).

We argue that the most important issue in determining statistical procedure should be the nature of the inference desired. If the analyst wishes to make inferences only about the effect-size parameters in the set of studies that are observed (or to a set of studies identical to the observed studies except for uncertainty associated with the sampling of participants into those studies), this is what we call a *conditional* inference. One might say that conditional inferences about the observed effect sizes are intended to be robust to the consequences of sampling error associated with sampling of participants (from the same populations) into studies. Strictly speaking, conditional inferences apply to *this* collection of studies and say nothing about other studies that may be done later, could have been done earlier, or may have already been done but are not included among the observed studies. Fixed-effects analysis procedures are appropriate for making conditional inferences.

In contrast, the analyst may wish to make a different kind of inference, one that embodies an explicit generalization beyond the observed studies. In this case, the studies that are observed are not the only studies that might be of interest. Indeed, one might say that the studies observed are of interest only because they reveal something about a putative population of studies that are the real object of inference. If the analyst wishes to make inferences about the parameters of a population of studies that is larger than the set of observed studies and that may not be strictly identical to them, we call this an *unconditional* inference. Random-effects analysis procedures are designed to facilitate unconditional inferences.

An Analogy to the ANOVA

To understand the fixed- and random-effects models in meta-analysis, it is helpful to place the problem in a context that is more familiar to many researchers: the ANOVA. Consider meta-analyses for which the data from different studies are directly comparable, so that the raw data from all the studies can be analyzed together. That is, assume that each study compares a treatment group with a control group and that all studies measure the outcome on the same scale and there is homogeneity of within-group variance across studies. This is not a case that arises in practice very often (because studies tend to use different outcome measures and have different sampling plans that lead to different variances), but it illustrates the ideas in a form that is easy to understand. We lose no generality in assuming that the common within-group variance is 1, which means that the raw mean difference in each study is an estimate of the standardized mean difference (Glass's effect size). It can be shown that the application of meta-analysis and the ANOVA yield the same results when applied to such a situation (Olkin & Sampson, 1998).

In this case, the data layout is a 2 (treatments) × k (studies) design, and we can apply ordinary two-factor ANOVA to analyze the data. The treatment factor is fixed, and the main effect of treatment corresponds to the average effect size (it is the weighted average of the study-specific treatment contrasts across studies). The Treatment × Studies interaction describes how much variation there is across studies in the study-specific treatment effects. The test of this interaction is a test that the study-specific treatment effect parameters are identical across studies.

Different tests for the fixed effect of treatment are appropriate depending on whether the studies factor is considered to be fixed or random. If the studies factor is treated as fixed, then the design is a fixed-effects design, and there is only one source of uncertainty, the within-group sampling error ϵ, and only one true variance component, the error variance σ_ϵ^2. The appropriate test uses only within-group variation (the mean square within groups) as the error term for testing the main effect of treatment. This corresponds to the intuition that the only source of uncertainty in inferences about the means of the observed studies is the sampling of participants into the studies. In this case, the inference is about the treatment-effect parameters in the particular studies included in the design.

If the studies factor is considered to be random, there are three sources of uncertainty: the within-group sampling error ϵ, the (random) effect of study β, and the Treatment × Study interaction $\alpha\beta$. Each of these sources of uncertainty has a corresponding variance component: σ_ϵ^2, σ_β^2, and $\sigma_{\alpha\beta}^2$. The fixed effects test for the main effect of treatment is incorrect; the correct error term is the Treatment × Studies interaction. This error term includes two components of variance: one due to within-study (within-group) sampling error σ_ϵ^2 and one due to the Study × Treatment interaction $\sigma_{\alpha\beta}^2$. Of course, the reason that this is the appropriate error term is that variance of the treatment contrast (and its associated mean square) also includes a component of variance due to the Treatment × Studies interaction $\sigma_{\alpha\beta}^2$. In this case, the inference about treatment-effect parameters is to the mean effect in a population of studies (a population of potential levels of the studies factor) from which the observed studies are a random sample.

The test for the Treatment × Studies interaction is the same in the two models, albeit with a slightly different interpretation. In the studies-fixed model, the test is that all the treatment-effect parameters in the observed studies are equal. In the studies-random model, the test is that the Treatment × Studies interaction variance component $\sigma_{\alpha\beta}^2$ is zero. This variance component describes the variance across all studies (in the putative population of studies) of the study-specific treatment-effect contrasts.

Inference to Other Studies

Fixed-effects models. In conditional (fixed-effects) models, inferences are, in the strictest sense, limited to the factor levels represented in the sample. Of course, conditional models are widely used in primary research, and the generalizations made from them by researchers are typically not constrained precisely to factor levels in the study. For example, generalizations about treatment effects in fixed-effects ANOVA are usually not constrained to apply only to the precise levels of treatment found in the experiment but are typically viewed as applying to similar treatments as well, even if they were not explicitly part of the experiment.

How are such inferences justified? Typically, they are justified on the basis of an a priori (extraempirical) decision that other levels (other treatments) are sufficiently like those in the sample that their behavior will be identical. The key point is that generalization to levels not present in the sample requires an assumption that the levels are similar to those in the sample—one not justified by a formal sampling argument.

Inference to studies not identical to those in the sample can be justified in meta-analysis by the same intellectual devices used to justify the corresponding inferences in primary research. Specifically, inferences may be justified if the studies are judged a priori to be sufficiently similar to those in the study sample. Note, however, that the inference process has two distinct parts. One part is the generalization from the study sample to a universe of identical studies, which is supported by a sampling theory rationale. The second part is the generalization from the universe of studies that are identical to the sample to a universe of sufficiently similar (but nonidentical) studies. This second part of the generalization is supported not by a sampling argument, but by an extrastatistical one.

Random-effects models. In random-effects models, inferences are not limited to studies represented in the sample. Instead, the inferences, for example, about the mean or variance of effect-size parameters, apply to the universe of studies from which the study sample was obtained. In effect, the warrant for generalization to other studies is through a classical sampling argument. Because the universe contains studies that differ in their characteristics and those differences find their way into the study sample by the process of random sampling, generalizations to the universe pertain to studies that are not identical to those in the study sample.

By using a sampling model of generalization, the random-effects model seems to avoid subjective difficulties that plague the fixed-effects model in generalizations to studies not identical to the study sample. That is, one does not have to ask, ''How similar is similar enough?'' Instead another question, ''Is this new study part of the universe from which the study sample was obtained?'' can be substituted. If study samples were obtained from well-defined sampling frames through overtly specified sampling schemes, this might be an easy question to answer. That, however, is virtually never the case in meta-analysis (and is unusual in other applications of random-effects models). The universe is usually rather ambiguously specified, and consequently, the ambiguity in generalization based on random-effects models is that it is difficult to know precisely what the universe is. In contrast, the universe is clear in fixed-effects models, but the ambiguity arises in deciding if a new study might be similar enough to the studies already contained in the study sample.

The random-effects model does provide the technical means to address an important problem that is not handled in the fixed-effects model, namely, the additional uncertainty introduced by the inference to studies that are not identical (except for the sample of people involved) to those in the study sample. Inference to (nonsampled) studies in the fixed-effects model occurs outside of the technical sampling theory framework, and hence, any uncertainty it contributes cannot be evaluated by technical means within the model. In contrast, the random-effects model does incorporate between-study variation into the sampling uncertainty used to compute tests and estimates.

Although the random-effects model has the advantage of incorporating inferences to a universe of studies exhibiting variation in their characteristics, the definition of the universe may be ambiguous. A tautological universe definition could be derived by using the sample of studies to define the universe as "a universe from which the study sample is representative." Such a population definition remains ambiguous; moreover, it may not be the universe definition desired for the use of the information produced by the synthesis. For example, if the study sample includes many studies of short-duration, high-intensity treatments but the likely practical applications usually involve low-intensity, long-duration treatments, the universe defined implicitly by the study sample may not be the universe most relevant to applications.

One potential solution to this problem might be to explicitly define a structured universe in terms of study characteristics and to consider the study sample as a stratified sample from this universe. Estimates of parameters describing this universe could be obtained by weighting each stratum appropriately. For example, if one half of the studies in the universe to which one wishes to generalize are long-duration studies but only one third of the study sample has this characteristic, the results of each long-duration study must be weighted twice as much as the short-duration studies.

Statistical Inference in Meta-Analysis

In this article, we assume that there are effect-size estimates from k independent studies. Denote the population effect size (effect-size parameter) in the ith study by θ_i and its estimate (the sample effect-size estimate) by T_i.

	Effect-	Effect-	Conditional
	size	size	variance
Study	parameter	estimate	of T given θ
1	θ_1	T_1	v_1
2	θ_2	T_2	v_2
.	.	.	.
.	.	.	.
.	.	.	.
k	θ_k	T_k	v_k.

We assume that the T_i is normally distributed about the corresponding θ_i with known variance v_i. That is, we assume that

$$T_i \sim N(\theta_i, v_i) \; i = 1, \ldots, k. \tag{1}$$

This assumption is very nearly exactly true for effect sizes such as the Fisher z-transformed correlation coefficient and standardized mean differences transformed by the Hedges–Olkin variance-stabilizing transformation (Hedges & Olkin, 1983). However, for effect sizes such as the untransformed standardized mean difference, correlation coefficient, or the log-odds ratio, the results are not exact but remain true as large-sample approximations.

Statistical Inference in Fixed-Effects Meta-Analysis

If a series of k studies can reasonably be expected to share a common effect size θ, or if we are interested in the mean of the effect sizes in the series of studies, it is natural to estimate θ by pooling estimates from each of the studies. If the sample sizes of the studies differ, then the estimates from the larger studies will usually be more precise than the estimates from the smaller studies. In this case, it is reasonable to give more weight to the more precise estimates when pooling. This leads to weighted estimators, and the weights that minimize the variance give weight inversely proportional to the variance in each study. This is intuitively clear in that smaller variance (i.e., more precision) should lead to a larger weight. The optimal weights are given by

$$w_i = 1/v_i. \tag{2}$$

Thus, the weighted mean that minimizes the variance can be written as

$$\overline{T}\cdot = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}. \qquad (3)$$

Note that $\overline{T}\cdot$ is also the maximum-likelihood estimator of θ under this model.

The sampling variance $v\cdot$ *of* $\overline{T}\cdot$ is simply the reciprocal of the sum of the weights, namely,

$$v\cdot = \frac{1}{\sum_{i=1}^{k} w_i}, \qquad (4)$$

and the standard error $SE(\overline{T}\cdot)$ of $\overline{T}\cdot$ is just the square root of $v\cdot$, that is, $SE(\overline{T}\cdot) = \sqrt{v\cdot}$. Because $T_1, \ldots, T_k$ is normally distributed, it follows that $\overline{T}\cdot$ also is normally distributed.

Tests and confidence intervals for the mean. If $T_1, \ldots, T_k$ estimates the same underlying effect size $\theta_1 = \ldots = \theta_k = \theta$, then $\overline{T}\cdot$ estimates θ and a 100(1 − α)% confidence interval for θ is given by

$$L = \overline{T}\cdot - z_{\alpha/2}\sqrt{v\cdot} \leq \theta \leq \overline{T}\cdot + z_{\alpha/2}\sqrt{v\cdot} = U, \quad (5)$$

where $z_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution and $v\cdot$ is the sampling variance of $\overline{T}\cdot$ given by Equation 4.

A 100α% significance test of the null hypothesis that $\theta = \theta_0$ could be obtained by means of the statistic

$$Z = (\overline{T}\cdot - \theta_0)/\sqrt{v\cdot}, \qquad (6)$$

which has the standard normal distribution when $\theta = \theta_0$. The one-sided test rejects the null hypothesis whenever $Z > z_\alpha$.

Testing for heterogeneity of effect-size parameters. Before pooling the estimates of effect size from a series of k studies, it is important to determine whether the studies can reasonably be described as sharing a common effect size. A statistical test for the homogeneity of population effect sizes is formally a test of the hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ versus the alternative that at least one of the effect sizes θ_i differs from the remainder.

An exact small-sample test of H_0 (which is also the likelihood ratio test of this hypothesis) is based on the statistic

$$Q = \sum_{i=1}^{k} w_i (T_i - \overline{T}\cdot)^2, \qquad (7)$$

where $\overline{T}\cdot$ is the weighted estimator of effect size given in Equation 3. The test statistic Q is the sum of squares of the T_i about the weighted mean $\overline{T}\cdot$ where the ith square is weighted by the reciprocal of the variance of T_i. Because $w_i = 1/v_i$ and $(T_i - \overline{T}\cdot)^2$ can be seen as a (crude) estimate of between-study variation, each term of Q can also be interpreted as a ratio of between-study to within-study variances, meaning that Q can be interpreted as a comparison of between- to within-study variance.

If all k studies have the same population effect size (i.e., if H_0 is true), then the test statistic Q has a chi-square distribution with $k - 1$ degrees of freedom. Therefore, if the obtained value of Q exceeds the 100(1 − α)% critical value of the chi-square distribution with $k - 1$ degrees of freedom, we reject the hypothesis that the θ_i are equal.

Example. The results of 14 studies of gender differences in field articulation ability were reported by Hyde (1981), who reported the effect-size estimate (as a standardized mean difference) and total sample size for each study. The data are listed in Table 1, where, for each study, column 2 gives the unbiased estimate of effect size corresponding to Hyde's standardized mean difference. To carry out the fixed-effects analysis, first compute the sampling variance v_i for each study. Hyde reported the total sample size for each study but not the sample sizes of each group. Consequently, we compute the variance here from the formula, on the basis of the assumption that the sample sizes of the two groups within a study are (approximately) equal (see Hedges, 1981), namely, $v_i = 4(1 + d_i^2/8)/N_i$, where N_i is the total sample size in the ith study.

The fixed-effects analysis depends on the sums of three variables: the weights ($w_i = 1/v_i$), the weights × the effect sizes ($w_i d_i$), and the weights × the effect sizes squared ($w_i d_i^2$). The analysis can be carried out by using a packaged computer program (e.g., SAS 6.2, 1996, or SPSS, 8.0, 1998) or a spreadsheet to compute these three variables and their sums. Column 3 of Table 1 gives the value of v_i for each study, and columns 4, 5, and 6 give the values of w_i, $w_i d_i$, and $w_i d_i^2$, along with their sums at the bottom of the columns.

The fixed-effects weighted mean effect size given in Equation 3 is just $\overline{T}\cdot = (118.050)/(216.700) = 0.545$, and its variance, given in Equation 4, is just $v\cdot = 1/216.700 = .004615$, which yields a standard error of $SE(\overline{T}\cdot) = \sqrt{(.004615)} = .068$. The 95% confidence interval for θ, given by Equation 5 is

$$0.412 = 0.545 - 1.96(.068) \leq \theta \leq 0.545 + 1.96(.068) = 0.678.$$

Table 1
Effect-Size Data From 14 Studies of Gender Difference in Field Articulation

Study	N	d	v	w	wd	wd^2	w^2	w^*	w^*d
1	60	0.76	.071	13.990	10.632	8.081	195.718	7.783	5.915
2	140	1.15	.033	30.035	34.540	39.721	902.093	11.075	12.736
3	30	0.48	.137	7.290	3.499	1.680	53.145	5.150	2.472
4	30	0.29	.135	7.422	2.152	0.624	55.086	5.216	1.513
5	30	0.65	.140	7.124	4.630	3.010	50.748	5.066	3.293
6	46	0.84	.095	10.568	8.877	7.457	111.681	6.595	5.540
7	40	0.70	.106	9.423	6.596	4.617	88.790	6.130	4.291
8	34	0.50	.121	8.242	4.121	2.061	67.938	5.608	2.804
9	76	0.18	.053	18.923	3.406	0.613	358.094	9.104	1.639
10	163	0.17	.025	40.603	6.903	1.173	1,648.630	12.251	2.083
11	97	0.77	.044	22.577	17.384	13.386	509.711	9.872	7.602
12	44	0.27	.092	10.901	2.943	0.795	118.825	6.723	1.815
13	78	0.40	.052	19.118	7.647	3.059	365.484	9.148	3.659
14	43	0.45	.095	10.485	4.718	2.123	109.927	6.563	2.953
Total				216.700	118.050	88.399	4,635.868	106.285	58.315

Note. These data are from Hyde (1981). v = conditional variance; w = weight, fixed effects model; w^* = weight, random-effects model.

The test of homogeneity of effect sizes is computed from Equation 7 as

$$Q = 88.399 - (118.050)^2/(216.700) = 24.090,$$

and because this value exceeds 22.36, the 95% critical value of the chi-square distribution with (14 − 1) = 13 degrees of freedom, we reject the hypothesis that the effect-size parameters are the same in all of the studies.

Statistical Inference in Random-Effects Meta-Analysis

In this section, we describe procedures for estimating the mean μ of the effect-size distribution underlying the results of a series of studies using an analysis based on the random-effects model. There are obvious similarities between estimating a common underlying effect size by taking the mean of the estimates and estimating the mean of the effect-size distribution. In both procedures, the pooled estimate is usually computed by taking the weighted mean across studies of the sample effect-size estimates, and it is not unusual for either of these estimates to be called the *average* effect size. However, note that the quantity to be estimated (the mean of the effect-size distribution) in random-effects models does not have exactly the same interpretation as the one to be estimated (the single or average underlying effect size) in fixed-effects models. In the case of random-effects models, for example, some individual effect-size parameters may be negative even though μ is positive. That corresponds to the substantive idea that some realizations of the treatment may actually be harmful even if the average effect of the treatment μ is beneficial.

The variance of estimates of effect size. In the fixed-effects model, the effect sizes θ_i are fixed, but unknown, constants. Under this assumption, the variance of T_i is simply v_i. In the random-effects model, the θ_i are not fixed but are themselves treated as random and have a distribution of their own. Therefore, it is necessary to distinguish between the variance of T_i assuming a fixed θ_i and the variance of T_i incorporating the variance of θ as well. The former is the *conditional* sampling variance of T_i, and the latter is the *unconditional* sampling variance of T_i.

It is convenient to decompose the observed effect-size estimate into fixed and random components

$$T_i = \theta_i + \epsilon_i = \mu + \xi_i + \epsilon_i, \tag{8}$$

where ϵ_i is a sampling error of T_i as an estimate of θ_i, and θ_i can itself be decomposed into the mean μ of the population from which the θs are sampled and the error ξ_i of θ_i as an estimate of μ. In this decomposition, only μ is fixed, and one can assume both ξ_i and the ϵ_is are random, with expected value zero. The variance of ϵ_i is v_i, the conditional sampling variance of T_i, which is known. The variance of the population from which $\xi_1, \ldots, \xi_k$ are sampled is τ^2. Equivalently, one might say that τ^2 is the variance of the population from which the study-specific effect parameters $\theta_1, \ldots, \theta_k$ are sampled. Frequently τ^2 is called the between-study variance component.

Because the effect size θ_i is a value obtained from a distribution of potential θ_i values, the unconditional sampling variance of T_i involves τ^2. A direct argument shows that this sampling variance is

$$v_i^* = v_i + \tau^2. \tag{9}$$

Methods of estimation for random-effects models have been suggested in different meta-analytic contexts by DerSimonian and Laird (1986), Hedges (1983), and Schmidt and Hunter (1977). They use the method of moments to estimate the between-study variance component and are analogous to the methods often used to estimate variance components in ANOVAs of balanced designs.

Estimating the between-studies variance component. Estimation of the between-studies variance component τ^2 uses the same principles as estimation of the variance components in the ANOVA. One estimate of τ^2 is

$$\hat{\tau}^2 = \begin{cases} \dfrac{Q-(k-1)}{c} & \text{if } Q \geqslant k-1 \\ 0 & \text{if } Q < k-1 \end{cases} \tag{10}$$

where c is given by

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} (w_i)^2}{\sum_{i=1}^{k} w_i}, \tag{11}$$

and w_i are the weights given in Equation 2 used in the fixed-effects analysis. Estimates of τ^2 are set to 0 when $Q - (k - 1)$ yields a negative value, because τ^2, by definition, cannot be negative.

If the within-study sampling-error variances $v_1, \ldots, v_k$, used to construct the weights w_i, are known exactly and the estimate is not truncated at 0, then the w_i are constants, and the estimate is unbiased, a result that does not depend on assumptions about the distribution of the random effects (or the conditional distribution of the effect sizes themselves). Inaccuracies in the estimation of the v_i (and hence the w_i) may lead to biases, although they are usually not substantial. The truncation of the estimate at zero is a more serious source of bias, although it improves the accuracy (reduces its mean square error about the true τ^2) of estimates of τ^2. This bias can be substantial when k is small but decreases rapidly when k becomes larger. Table 2 gives the bias of $\hat{\tau}^2$ when $v_1 = \ldots = v_k = v$ for values of k and τ^2. The table shows that the absolute bias of $\hat{\tau}^2$ can be as much as .2 to .3 for $k = 3$ and $\tau^2 = .33v$, leading to relative biases that are well over 50%. This result underscores the fact that estimates of τ^2 computed from only a few studies should be treated with caution. For $k > 20$, the biases are much smaller, and relative biases are only a few percent.

Note that other methods of estimation of variance components are available. For example, maximum-likelihood estimation under either a restricted model (first estimating the mean and then estimating the variance component conditional on the estimate of the mean) or unrestricted model (estimating the mean and the between-studies variance component simultaneously) is used in other problems and can be used in meta-analysis (see, e.g., Raudenbush & Bryk, 1985). In the case of equal v_i, which we examine, the method of moments is identical to restricted maximum-likelihood estimation. These methods are iterative and therefore more complex to implement in general, usually requiring specialized computer programs.

Testing the significance of the effect-size variance component. The test that $\tau^2 = 0$ in the random-effects model is the same as the test of homogeneity in the fixed-effects model, using the Q statistic. The reason is that if $\tau^2 = 0$, then $\theta_1 = \theta_2 = \ldots = \theta_k = \mu$; thus, the effect-size parameters are fixed, but unknown, constants. This is analogous to the situation with the F tests in the one-way random- and fixed-effects ANOVAs. In the ANOVA, the null distributions of the test statistics are identical, but the nonnull distributions of the F ratios differ. Similarly, although the null distributions of the Q statistics are identical in fixed- and random-effect-size models, the nonnull distributions of Q differ under the two models.

Estimating the mean effect size. The logic of using weighting is the same in random-effects procedures as it is in fixed-effects procedures, but the choice of weights differs somewhat because random-effects models include in their definition of *variance* a component of variance τ^2 associated with between-study differences in effect parameters, which fixed-effects models do not. That is, the total variance v_i^* for the ith effect-size estimate T_i is defined by $v_i^* = \tau^2 + v_i$. Because the additional component of variance is the same for all studies, it both increases the total variance of each effect size estimate and tends to make the total variances of the studies (the v_i^*) more equal than the sampling-error variances (the v_i).

Because the true value of τ^2 is rarely known, we usually substitute an estimate of this variance compo-

Table 2
Bias $E(\hat{\tau}^2) - \tau^2$ and Relative Bias $[E(\hat{\tau}^2) - \tau^2]/\tau^2$ in the Estimator $\hat{\tau}^2$ of τ^2 Based on k Studies

	Bias				Relative Bias		
k	$\tau^2 = 0$	$\tau^2 = .33$	$\tau^2 = .67$	$\tau^2 = 1.00$	$\tau^2 = .33$	$\tau^2 = .67$	$\tau^2 = 1.00$
2	0.484	0.429	0.389	0.358	1.286	0.583	0.358
3	0.368	0.296	0.248	0.213	0.889	0.372	0.213
4	0.308	0.229	0.179	0.144	0.688	0.268	0.144
5	0.271	0.187	0.137	0.104	0.562	0.205	0.104
6	0.244	0.158	0.108	0.078	0.474	0.163	0.078
7	0.224	0.137	0.088	0.060	0.410	0.132	0.060
8	0.208	0.120	0.073	0.047	0.360	0.110	0.047
9	0.195	0.106	0.061	0.038	0.319	0.092	0.038
10	0.185	0.095	0.052	0.030	0.286	0.078	0.030
20	0.128	0.043	0.014	0.005	0.128	0.022	0.005
30	0.104	0.024	0.005	0.001	0.072	0.008	0.001
40	0.090	0.015	0.002	0.000	0.044	0.003	0.000
50	0.080	0.010	0.001	0.000	0.029	0.001	0.000
60	0.073	0.007	0.000	0.000	0.020	0.001	0.000
70	0.068	0.005	0.000	0.000	0.014	0.000	0.000
80	0.063	0.003	0.000	0.000	0.010	0.000	0.000
90	0.060	0.002	0.000	0.000	0.007	0.000	0.000
100	0.057	0.002	0.000	0.000	0.005	0.000	0.000

nent such as that given in Equation 10 into Equation 9 in place of τ^2 (see, e.g., DerSimonian & Laird, 1986; Hedges, 1983; Hedges & Olkin, 1985). This yields

$$\overline{T}.^* = \frac{\sum_{i=1}^{k} w_i^* \overline{T}_i^*}{\sum_{i=1}^{k} w_i^*}, \qquad (12)$$

where the weight w_i^* is an estimated optimal weight that is the reciprocal of an estimate of the total variance of $\overline{T}_i$ given by

$$w_i^* = 1/(v_i^*) = 1/(v_i + \hat{\tau}^2). \qquad (13)$$

Here, we use the * to distinguish the weights, means, and variances in the random-effects procedure from the corresponding quantities in the fixed-effects procedure.

The sampling variance $v.^*$ of the random-effects estimate (of the mean of the effect-size distribution) $\overline{T}.^*$ is given by the reciprocal of the sum of the random-effects weights, that is,

$$v.^* = \frac{1}{\sum_{i=1}^{k} w_i^*}. \qquad (14)$$

The standard error $SE(\overline{T}.^*)$ of the mean effect estimate $\overline{T}.^*$ is just the square root of its sampling variance, that is, $SE(\overline{T}.^*) = \sqrt{v.^*}$. Note that whenever the between-studies variance component (estimate) $\hat{\tau}^2 > 0$, the standard error $\sqrt{v.^*}$ of the mean estimated using the random-effects procedure will be larger than $\sqrt{v.}$, the standard error of the mean estimated using the fixed-effects procedure. If $\hat{\tau}^2 = 0$, the standard errors (and the mean estimates) of the random- and fixed-effects procedures will be identical.

Note that some writers who advocate random-effects procedures (e.g., Hunter & Schmidt, 1990, p. 147) advocate the use of suboptimal weights that correspond to the fixed-effects weights, presumably because they assume that τ^2 is small. This will not make a great deal of difference if τ^2 is indeed small, but assuming that $\tau^2 = 0$ when it is not will lead to an underestimate of the variance $v.^*$.

Although it is possible to obtain very precise estimates of θ_i if the sample sizes are reasonably large in each study, the precision of the estimate of τ^2 depends primarily on the number k of studies. Therefore, if the number of studies is small, the estimates of the weights may be fairly imprecise even though the sample size in each study is quite large.

Tests and confidence intervals for the mean. If the random effects are approximately normally distributed, the weighted mean $\overline{T}.^*$ is approximately normally distributed about the mean effect-size parameter μ that it estimates. As in the fixed-effects case, the fact that this mean is normally distributed with the variance given in Equation 14 leads to straightforward procedures for constructing tests and confidence intervals. An approximate $100(1 - \alpha)\%$ confidence interval for the mean effect μ is given by

$$L^* = \overline{T}.^* - z_{\alpha/2}\sqrt{v.^*} \leqslant \mu \leqslant \overline{T}.^* + z_{\alpha/2}\sqrt{v.^*} = U^*, \quad (15)$$

where $z_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution (e.g., $z_{\alpha/2} = 1.96$ for $\alpha = .05$ and 95% confidence intervals) and $v.^*$ is the variance of $\overline{T}.^*$ given in Equation 14.

Significance tests corresponding to the confidence intervals also can be constructed. An approximate test of whether the mean effect μ differs from a predefined constant μ_0 (e.g., to test if $\mu - \mu_0 = 0$) by testing the null hypothesis H_0: $\mu = \mu_0$, uses the statistic

$$Z^* = \frac{\overline{T}* - \mu_0}{(v.^*)^{1/2}}.$$

The one-sided test consists of rejecting H_0 at level α (that is, decide that the effect parameter differs from μ_0) if the value of Z^* exceeds the $100\alpha\%$ critical value of the standard normal distribution. That is, reject H_0 if $Z^* > z_\alpha$. For example, for a one-sided test that $\mu = 0$ at $\alpha = .05$ level of significance, reject the null hypothesis if the value of Z^* exceeds 1.645.

Example. Return to the studies of gender differences in field articulation ability discussed above and whose results are reported in Table 1. The random-effects analysis depends on the sums of six variables: the fixed-effects weights ($w_i = 1/v_i$), the fixed-effects weights × the effect sizes (w_id_i), the fixed-effects weights × the effect sizes squared ($w_id_i^2$), the fixed-effects weights squared (w_i^2), the random-effects weights (w_i*), and the random-effects weights × the effect sizes (w_i*d_i). The sums of the first four of these variables are needed to compute the variance component estimate, which in turn is used to compute the random-effects weights in the last two variables. The analysis can be carried out by using a packaged computer program (e.g., SAS 6.12, 1996, or SPSS 8.0, 1998) or a spreadsheet, to compute these six variables and their sums. In Table 1, columns 4–9 give the values of w_i, w_id_i, $w_id_i^2$, w_i^2, w_i*, and w_i*d_i, for each study, along with their sums at the bottom of the columns.

The first step in the random-effects analysis is to compute the between-studies variance component estimate. The homogeneity test statistic Q was computed in the fixed-effects analysis as $Q = 24.090$. The constant c given in Equation 11 is computed from the sums of w_i and w_i^2 as

$$c = 216.700 - (4{,}635.868)/(216.700) = 195.307,$$

and the variance component estimate itself is computed from Equation 10 as

$$\hat{\tau}^2 = [24.090 - (14 - 1)]/195.307 = 0.057.$$

Given that $\hat{\tau}^2 = 0.057$, the random-effects weights are computed, using Equation 13, as $w_i^* = 1/(v_i + 0.057)$. Note that in a spreadsheet, this computation can easily be automated, but in a computer program package such as SAS 6.12, 1996, or SPSS 8.0, 1998, the random-effects analysis will require two passes through the data: one to obtain the sums necessary to compute $\hat{\tau}^2$ and a second pass to compute the sum of the w_i* and w_i*d_i.

The random-effects weighted mean effect size given in Equation 12 is just $\overline{T}.^* = (58.315)/(106.285) = 0.549$, and its variance, given in Equation 14, is just $v.^* = 1/106.285 = 0.009408$, which yields a standard error of $SE(\overline{T}.^*) = \sqrt{(0.009408)} = .097$. The 95% confidence interval for μ, given by Equation 15, is

$$0.359 = 0.549 - 1.96(.097) \leqslant \mu \leqslant 0.549 + 1.96(.097) = 0.739.$$

Comparing the results of the random-effects analysis with the fixed-effects analysis of the same data given in the previous example reveals that the weighted means computed in the two analyses are almost identical ($\overline{T}.^* = 0.549$ vs. $\overline{T}. = .545$), but the standard error computed in the random-effects analysis is substantially larger than that in the fixed-effects analysis, $SE(\overline{T}.^*) = 0.097$ versus $SE(\overline{T}.) = 0.068$. The reason for the difference in standard errors is the substantial between-study heterogeneity in the effect sizes. The between-study variance component estimate (0.057) is about two thirds as large as the average (0.086) of the sampling error variances. Consequently, when this substantial additional component of variance is included as part of the sampling uncertainty of the mean in the random-effects model, the standard error of the mean sharply increases.

Conditional Choice of Random-Effects Procedures

We have treated the choice between fixed- and random-effects procedures as an a priori one based primarily on conceptual issues, not on the outcomes of the analysis. However, the emphasis on homogeneity as the criterion for choosing between fixed- and random-effects procedures has led some to use the statistical test of heterogeneity as the sole criterion for choice among statistical procedures. That is, a preliminary test is conducted to determine whether $\tau^2 > 0$; if the null hypothesis that $\tau^2 = 0$ is rejected, a random-effects procedure is used, and if the hypothesis $\tau^2 = 0$ is not rejected, then a fixed-effects procedure is used. The properties of such a conditional procedure were first analyzed by Chang (1992), who investigated them by means of simulation. We call this procedure the *conditionally random-effects* procedure, to emphasize that the choice of random (vs. fixed) effects is conditional on the outcome of the test that $\tau^2 > 0$. This procedure has appeal because it is simple and offers an alternative to either the fixed- or random-effects procedures. It is also equivalent to the practice of the many meta-analysts who use the test for homogeneity of effect size, to determine whether effects differ across studies, and then use random-effects analysis procedures if homogeneity is rejected, and fixed-effects analysis procedures if homogeneity is not rejected.

Comparing Inference Procedures

To compare the performance of inference procedures, it is essential to clarify the inferences being drawn by them. Specifically, we must clarify the difference between *conditional* inferences drawn about the mean of the specific set of effect parameters in the set of studies analyzed and *unconditional* inferences about the mean of the population of effect parameters from which the observed study parameters are a random sample.

It is important to distinguish between the underlying statistical *model,* which is determined by the inferences desired, and the statistical *procedure* used, which is a choice made by the analyst to accomplish the inferential purpose. If the analyst chooses to make conditional inferences (by conditioning on the studies in the data set), the statistical model has been determined because the effect parameters are treated as fixed for inference. If the analyst chooses to make unconditional inferences, the statistical model treats the effect sizes as a sample (even if no real sampling has been done), and thus, they are treated as random effects.

We analyze the performance of fixed, random, and conditionally random statistical procedures under both conditional and unconditional inference models. First, we examine the rejection rates of tests on means and the probability content of confidence intervals when conditional inferences are desired. Then we consider the rejection rates of tests on means and the probability content of confidence intervals when unconditional inferences are desired.

Analytic comparisons among the three inference methods are not easy in general, but substantial insight can be obtained by examining the case in which the effect-size parameters are normally distributed and all the conditional variances are equal, that is, $v_1 = \ldots = v_k = v$ (which usually is equivalent to the condition that the within-study sample sizes are equal). In this case, the weights in all three inference procedures are equal across studies. Consequently, the mean effect sizes under each procedure are simply the unweighted means of the T_i. The variances computed using the fixed- and random-effects procedures will differ, however, in that the fixed-effects estimate of the variance will be $v. = v/k$ and the random-effects estimate of the variance is $v.^* = (v + \hat{\tau}^2)/k$.

The random effects and conditionally random-effects confidence intervals and tests on the mean effect size involve the use of the variance component estimate $\hat{\tau}^2$ in place of τ^2 to construct the weights used in computing the mean $\overline{T}.^*$ and its variance $v.^*$. If the number of studies is small, the estimate of the variance component will not be very accurate. It is reasonable to ask therefore how accurate these inferences might be when the number of studies is small.

Conditional Inferences

When conditional inferences are desired, the analyst conditions on the study characteristics, so that the effect parameters are fixed but unknown constants for inference. Statistical inferences are about the mean of the effect parameters in the experiments analyzed. Specifically, when conditional inferences are desired, the test on the mean effect size is a test of the null hypothesis H_0: $\overline{\theta} = \theta_0$, where θ_0 is some predefined constant (often 0) and the relevant confidence intervals are confidence intervals for $\overline{\theta}$, the mean of the k effect-size parameters $\theta_1, \ldots, \theta_k$ in the k studies being analyzed. Conditionally, the sampling distribution of the mean effect-size estimate is normal with a variance v/k.

In the fixed-effects analysis, the effect parameters are treated as fixed, and the variance of the mean is exactly correct. Therefore, the rejection rates and probability content of the confidence intervals based on the fixed-effects procedures can be calculated directly and correspond exactly to the nominal values. Because the variance of $\overline{T}_{\cdot}^{*}$ computed using the random- and conditionally random-effects procedures includes a contribution from the variance component estimate $\hat{\tau}^2$, it will be too large. Therefore, the test on the mean effect size will have a rejection rate that is less than the nominal significance level when H_0 is true, and the probability content of confidence intervals will be larger than the nominal values.

The variance of the mean effect-size estimate using random- and conditionally random-effects procedures depends on the variance component estimate $\hat{\tau}^2$. Therefore, both the probability content of the confidence intervals and the rejection rates of the test statistic can be evaluated by conditioning on $\hat{\tau}^2$ and averaging over values of $\hat{\tau}^2$ (weighting according to their probability density). The details are given in the Appendix. Table 3 contains the probability content of (nominally) 95% confidence intervals for $\overline{\theta}$, based on the three procedures, for sample sizes ranging from 2 studies to 100 studies and degrees of heterogeneity ranging from 0% to 100% of the sampling-error variance. These values of heterogeneity seem plausible, as $\tau^2 = 0$ corresponds to complete homogeneity, whereas $\tau^2 = .33v$ corresponds to the situation in which 75% of the total variance among effect sizes is sampling error, and $\tau^2 = v$ corresponds to the situation in which 50% of the total variance is sampling error, which are consistent with the range of values found in the survey conducted by Schmidt (1992).

As expected, the random- and conditionally random-effects procedures lead to confidence intervals that have larger than nominal probability content. In other words, they produce confidence intervals that are too wide. The extent of the exaggeration in width depends on the number of studies and the amount of between-study heterogeneity, but it can be profound if both are large.

Table 4 contains the rejection rate of the (nominal $\alpha = .05$) one-tailed test of the null hypothesis that $\overline{\theta} = 0$ for each of the inference procedures. The amounts of heterogeneity (variance between effect-size parameters) are the same as in Table 3. It is evident that the rejection rates are exactly nominal for the fixed-effects procedures but lower than nominal for the random- and conditionally random-effects procedures. The extent to which rejection rates are too low for random- and conditionally random-effects procedures depends on heterogeneity and the number of studies, but the departure from nominal can be profound when both are large. Table 5 contains the rejection rates for the same test when the null hypothesis is false and $\overline{\theta} = .25\sqrt{v}$, which gives some indication of power. In general, the rejection rate is highest for the fixed-effects test, followed by that of the conditionally random-effects test and then the random-effects test.

Table 3
Probability Content of 95% Confidence Intervals Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Conditional Inferences

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.00$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.950	.952	.962	.950	.956	.969	.950	.961	.975	.950	.965	.980
3	.950	.952	.963	.950	.957	.970	.950	.961	.977	.950	.966	.982
4	.950	.952	.963	.950	.957	.971	.950	.962	.978	.950	.968	.983
5	.950	.952	.962	.950	.957	.971	.950	.963	.979	.950	.970	.985
6	.950	.952	.962	.950	.958	.971	.950	.964	.979	.950	.971	.985
7	.950	.952	.962	.950	.958	.972	.950	.965	.980	.950	.973	.986
8	.950	.952	.962	.950	.958	.972	.950	.966	.980	.950	.974	.987
9	.950	.952	.961	.950	.959	.972	.950	.967	.981	.950	.976	.987
10	.950	.952	.961	.950	.959	.972	.950	.968	.981	.950	.977	.988
20	.950	.952	.959	.950	.961	.973	.950	.974	.984	.950	.986	.991
30	.950	.952	.958	.950	.963	.973	.950	.979	.985	.950	.990	.992
40	.950	.952	.957	.950	.965	.973	.950	.982	.986	.950	.992	.993
50	.950	.952	.957	.950	.966	.974	.950	.984	.986	.950	.993	.993
60	.950	.952	.956	.950	.967	.974	.950	.985	.987	.950	.993	.993
70	.950	.951	.956	.950	.968	.974	.950	.986	.987	.950	.993	.993
80	.950	.951	.956	.950	.969	.974	.950	.986	.987	.950	.994	.994
90	.950	.951	.955	.950	.970	.975	.950	.987	.987	.950	.994	.994
100	.950	.951	.955	.950	.970	.975	.950	.987	.987	.950	.994	.994

Table 4
Rejection Rates for Nominal $\alpha = .05$, One-Tailed Tests of the Hypothesis That $\theta = 0$ Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Conditional Inferences

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.00$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.050	.048	.039	.050	.044	.032	.050	.039	.026	.050	.035	.022
3	.050	.048	.039	.050	.043	.031	.050	.039	.025	.050	.034	.020
4	.050	.048	.039	.050	.043	.031	.050	.038	.024	.050	.033	.019
5	.050	.048	.039	.050	.043	.031	.050	.037	.024	.050	.031	.018
6	.050	.048	.040	.050	.043	.031	.050	.036	.023	.050	.029	.017
7	.050	.048	.040	.050	.042	.031	.050	.035	.023	.050	.028	.017
8	.050	.048	.040	.050	.042	.031	.050	.035	.023	.050	.027	.016
9	.050	.048	.040	.050	.042	.031	.050	.034	.022	.050	.026	.016
10	.050	.048	.041	.050	.042	.031	.050	.033	.022	.050	.025	.015
20	.050	.048	.042	.050	.040	.031	.050	.028	.020	.050	.017	.013
30	.050	.048	.043	.050	.039	.031	.050	.025	.020	.050	.014	.012
40	.050	.049	.044	.050	.038	.031	.050	.022	.019	.050	.012	.012
50	.050	.049	.045	.050	.037	.030	.050	.021	.019	.050	.012	.011
60	.050	.049	.045	.050	.036	.030	.050	.020	.018	.050	.011	.011
70	.050	.049	.045	.050	.035	.030	.050	.019	.018	.050	.011	.011
80	.050	.049	.046	.050	.034	.030	.050	.018	.018	.050	.011	.011
90	.050	.049	.046	.050	.034	.030	.050	.018	.018	.050	.011	.011
100	.050	.049	.046	.050	.033	.030	.050	.018	.018	.050	.011	.011

Table 5
Rejection Rates for Nominal $\alpha = 0.05$, One-Tailed Tests of the Hypothesis that $\theta = 0$ When the True $\theta = .25\sqrt{v}$, Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Conditional Inferences

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.00$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.098	.097	.078	.098	.086	.064	.098	.078	.053	.098	.070	.044
3	.113	.111	.089	.113	.098	.073	.113	.088	.059	.113	.077	.047
4	.126	.123	.101	.126	.109	.082	.126	.096	.065	.126	.083	.051
5	.139	.135	.112	.139	.120	.091	.139	.104	.072	.139	.087	.055
6	.151	.147	.124	.151	.130	.100	.151	.111	.078	.151	.092	.060
7	.163	.159	.135	.163	.140	.108	.163	.118	.084	.163	.095	.064
8	.174	.170	.146	.174	.149	.117	.174	.125	.090	.174	.099	.068
9	.185	.181	.157	.185	.159	.126	.185	.131	.097	.185	.102	.072
10	.196	.192	.167	.196	.168	.134	.196	.138	.103	.196	.106	.076
20	.299	.293	.269	.299	.254	.218	.299	.196	.166	.299	.141	.123
30	.391	.385	.363	.391	.333	.299	.391	.254	.232	.391	.185	.176
40	.475	.468	.448	.475	.407	.375	.475	.313	.298	.475	.236	.233
50	.549	.543	.524	.549	.475	.447	.549	.373	.364	.549	.293	.291
60	.615	.609	.592	.615	.538	.514	.615	.433	.427	.615	.350	.350
70	.672	.668	.653	.672	.595	.575	.672	.491	.488	.672	.408	.408
80	.723	.718	.705	.723	.647	.631	.723	.547	.544	.723	.464	.464
90	.766	.763	.751	.766	.693	.680	.766	.598	.597	.766	.517	.517
100	.804	.800	.791	.804	.735	.725	.804	.646	.645	.804	.568	.568

Unconditional Inferences

When unconditional inferences are desired, the analyst treats the effect parameters as a sample from a population and estimates the mean and variance of that population. Statistical inferences are about the mean (and possibly the variance) of the population from which the study effect sizes were sampled. Specifically, when unconditional inferences are desired, the test on the mean effect size is a test of the null hypothesis H_0: $\mu = \mu_0$, where μ_0 is some predefined constant (often 0) and the relevant confidence intervals are confidence intervals for μ, the mean of the population from which the k effect-size parameters $\theta_1, \ldots, \theta_k$ in the k studies being analyzed were sampled. Unconditionally, the sampling distribution of the mean is normal with a variance $(v + \tau^2)/k$.

In the fixed-effects analysis, the effect parameters are treated as fixed, and the variance of the mean effect-size estimate is underestimated as v/k. The rejection rates and probability content of the confidence intervals based on the fixed-effects procedures can be calculated directly but do not correspond to the nominal values. The test on the mean effect size will have a rejection rate that is larger than the nominal significance level when H_0 is true, and the probability content of confidence intervals will be less than the nominal values.

The random- and conditionally random-effects procedures should produce more accurate estimates of the variance of the mean effect. However, they will not be exactly correct because both involve using the variance component estimate $\hat{\tau}^2$ in place of the true value of the variance component τ^2. Both the probability content of the confidence intervals and the rejection rates of the test can be evaluated by conditioning on $\hat{\tau}^2$ and averaging over values of $\hat{\tau}^2$ (weighting according to their probability density). The details are given in the Appendix.

Table 6 contains the probability content of (nominally) 95% confidence intervals for μ based on the three procedures, for sample sizes ranging from 2 studies to 100 studies and degrees of heterogeneity ranging from 0% to 100% of the sampling error variance. As expected, the fixed-effects procedure yields confidence intervals that have exactly the nominal probability content when $\tau^2 = 0$, but when $\tau^2 > 0$, they have lower than nominal probability content (are too narrow). Random- and conditionally random-effects procedures lead to confidence intervals that have slightly larger than nominal probability content when $\tau^2 = 0$ (because they overestimate τ^2 in this case) and less than nominal content (are too narrow) when $\tau^2 > 0$. The extent of the underestimation in width is greater when the number of studies is small and the

Table 6
Probability Content of 95% Confidence Intervals Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Unconditional Inferences

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.00$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.950	.951	.962	.911	.913	.936	.871	.879	.913	.834	.848	.893
3	.950	.951	.963	.911	.915	.940	.871	.883	.921	.834	.857	.905
4	.950	.951	.963	.911	.916	.941	.871	.887	.924	.834	.865	.912
5	.950	.951	.962	.911	.917	.942	.871	.890	.927	.834	.871	.916
6	.950	.951	.962	.911	.918	.943	.871	.893	.929	.834	.877	.919
7	.950	.951	.962	.911	.919	.943	.871	.896	.930	.834	.882	.922
8	.950	.951	.962	.911	.920	.943	.871	.898	.931	.834	.886	.924
9	.950	.951	.961	.911	.920	.943	.871	.900	.932	.834	.891	.926
10	.950	.951	.961	.911	.921	.944	.871	.902	.933	.834	.894	.927
20	.950	.951	.959	.911	.925	.945	.871	.917	.939	.834	.919	.937
30	.950	.951	.958	.911	.929	.945	.871	.927	.942	.834	.932	.941
40	.950	.951	.957	.911	.931	.946	.871	.933	.943	.834	.939	.943
50	.950	.951	.957	.911	.933	.946	.871	.938	.945	.834	.942	.944
60	.950	.951	.956	.911	.935	.947	.871	.941	.945	.834	.944	.945
70	.950	.951	.956	.911	.937	.947	.871	.943	.946	.834	.945	.946
80	.950	.951	.956	.911	.938	.947	.871	.944	.946	.834	.946	.946
90	.950	.951	.955	.911	.939	.947	.871	.945	.947	.834	.947	.947
100	.950	.951	.955	.911	.941	.948	.871	.946	.947	.834	.947	.947

heterogeneity is large, and it can be substantial in extreme cases. However, the fixed-effects procedures are always further from nominal than either random- or conditionally random-effects procedures.

Table 7 contains the rejection rate of the (nominal $\alpha = .05$) one-tailed test of the null hypothesis that $\mu = 0$ for each of the inference procedures. The amounts of heterogeneity (variance between) effect-size parameters are the same as in Table 6. It is evident that the rejection rates for the fixed-effects procedures are exactly nominal when $\tau^2 = 0$, but those of the random- and conditionally random-effects procedures are lower than nominal. The rejection rates for all procedures are higher than nominal when $\tau^2 > 0$, but those of the random- and conditionally random-effects procedures are closer to nominal than those of the fixed-effects procedures. When $\tau^2 > 0$, the rejection rate converges to the nominal as the number of studies increases and is reasonably close to nominal if $k > 20$. Table 8 contains the rejection rates for the same test when the null hypothesis is false and $\mu = .25\sqrt{v}$, which gives some indication of power. In general, the rejection rate is highest for the fixed-effects test, followed by that of the conditionally random-effects test and then the random-effects test. However, in comparing the power of these tests, it is important to remember that the null ($\mu = 0$) rejection rate of the fixed-effects procedure is also higher than nominal when $\tau^2 > 0$.

Effects of Unequal Variances

The calculations in this article were based on equal within-study conditional variances ($v_1 = \ldots = v_k$). These calculations provide an indication of the magnitude of the differences in performance among procedures that might be expected. It is reasonable to ask how these results would generalize to more realistic cases of unequal conditional variances. Fortunately, qualitative arguments suggest that the differences in performance among procedures will remain similar even when the conditional variances are unequal.

With unequal variances, the rejection rates under the null hypothesis and probability content of confidence intervals for fixed-effects analyses would still be exactly correct in the case of conditional inferences. Random- and (to a lesser extent) conditionally random-effects procedures overestimate the variance of the mean effect size under the conditional inference model, although the exact amount will depend on the configuration of conditional variances. Thus, these procedures will have smaller than nominal rejection rates under the null hypothesis and confidence intervals that have higher than nominal probability content.

Table 7
Rejection Rates for Nominal $\alpha = .05$, One-Tailed Tests of the Hypothesis That $\mu = 0$, Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Unconditional Inferences

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.00$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.050	.049	.039	.077	.075	.057	.102	.095	.071	.122	.112	.082
3	.050	.049	.039	.077	.073	.054	.102	.092	.066	.122	.106	.075
4	.050	.049	.039	.077	.072	.054	.102	.089	.064	.122	.101	.071
5	.050	.049	.039	.077	.071	.053	.102	.087	.063	.122	.097	.069
6	.050	.049	.040	.077	.071	.053	.102	.085	.062	.122	.093	.067
7	.050	.049	.040	.077	.070	.053	.102	.083	.061	.122	.090	.066
8	.050	.049	.040	.077	.070	.053	.102	.082	.060	.122	.087	.064
9	.050	.049	.040	.077	.069	.053	.102	.080	.060	.122	.085	.063
10	.050	.049	.041	.077	.069	.053	.102	.079	.059	.122	.082	.063
20	.050	.049	.042	.077	.066	.053	.102	.070	.056	.122	.068	.057
30	.050	.049	.043	.077	.064	.052	.102	.064	.055	.122	.060	.055
40	.050	.049	.044	.077	.062	.052	.102	.060	.054	.122	.056	.054
50	.050	.049	.045	.077	.061	.052	.102	.057	.053	.122	.054	.053
60	.050	.049	.045	.077	.060	.052	.102	.056	.053	.122	.053	.053
70	.050	.049	.045	.077	.058	.052	.102	.054	.052	.122	.053	.052
80	.050	.049	.046	.077	.058	.052	.102	.053	.052	.122	.052	.052
90	.050	.049	.046	.077	.057	.051	.102	.053	.052	.122	.052	.052
100	.050	.049	.046	.077	.056	.051	.102	.052	.052	.122	.052	.052

Table 8
Rejection Rates for Nominal $\alpha = .05$, One-Tailed Tests of the Hypothesis That $\mu = 0$ When the True $\mu = .25\sqrt{v}$, Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Unconditional Inferences

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.00$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.098	.093	.076	.132	.120	.096	.159	.138	.111	.181	.151	.121
3	.113	.111	.089	.147	.139	.106	.174	.158	.117	.196	.169	.124
4	.126	.123	.101	.160	.150	.116	.188	.166	.124	.209	.173	.124
5	.139	.135	.112	.173	.161	.125	.200	.173	.131	.221	.177	.128
6	.151	.147	.124	.185	.171	.134	.212	.180	.138	.233	.180	.133
7	.163	.159	.135	.197	.180	.144	.223	.186	.145	.243	.184	.138
8	.174	.170	.146	.208	.190	.152	.234	.193	.151	.254	.187	.143
9	.185	.181	.157	.219	.199	.161	.244	.199	.158	.263	.191	.153
10	.196	.192	.167	.229	.208	.170	.254	.205	.165	.273	.194	.158
20	.299	.293	.269	.324	.288	.251	.342	.260	.228	.355	.230	.209
30	.391	.385	.363	.406	.359	.325	.416	.310	.287	.423	.269	.258
40	.475	.468	.448	.478	.424	.393	.480	.359	.343	.482	.311	.306
50	.549	.543	.524	.542	.483	.456	.538	.407	.396	.535	.354	.351
60	.615	.609	.592	.600	.537	.514	.589	.453	.446	.582	.396	.395
70	.672	.668	.653	.651	.586	.567	.635	.497	.493	.624	.438	.437
80	.723	.718	.705	.696	.631	.615	.676	.539	.536	.662	.477	.477
90	.766	.763	.751	.736	.672	.659	.713	.579	.577	.696	.515	.515
100	.804	.800	.791	.771	.709	.698	.746	.617	.615	.727	.550	.550

In the case of unconditional inferences, the rejection rate of fixed-effects tests under the null hypothesis would be larger than nominal and the probability content of confidence intervals smaller than nominal whenever $\tau^2 > 0$. This is because fixed-effects procedures would underestimate (to an extent that depends on the configuration of conditional variances) the actual unconditional variance of the mean effect-size estimate. Random- and (and to a lesser extent) conditionally random-effects procedures would provide rejection rates and confidence intervals whose probability content were closer to nominal but would not be exactly nominal because they involved substituting an estimate of τ^2 for the exact value of τ^2 in computing weights and the variance of the mean effect size. However, as the number of studies increased, the performance of both conditionally random and random-effects procedures would converge to nominal.

Conclusion

The selection of a statistical procedure in meta-analysis (and elsewhere) should be determined by the inferences one wishes to make. Once the inference goals are clear, a statistical procedure appropriate for those goals should be chosen. If conditional inferences (i.e., inferences about the parameters in the collection of studies observed) are desired, then fixed-effects procedures should be used. Heterogeneity of effects should not necessarily be a reason to abandon fixed-effects analyses. However, substantial heterogeneity may suggest that explanatory analyses are desirable to gain an understanding of that heterogeneity. Using random-effects or conditionally random-effects analyses to make conditional inferences will result in less powerful tests of significance (tests that have a higher actual significance level than the nominal) and confidence intervals that are too wide.

If unconditional inferences (i.e., inferences about the population from which the observed studies are sampled) are desired, then either conditionally random- or random-effects procedures should be used. Unless there is almost perfect homogeneity of effects ($\tau^2 = 0$ or nearly so), fixed-effects tests will reject more often than expected (have a higher actual significance level than nominal). Random-effects tests and confidence intervals will not give results that are exactly nominal either, particularly if the number of studies is small, but their performance will be closer to nominal than fixed-effects tests. If the number of studies is very small (say less than five), random-effects tests should be regarded as only approximate.

Conditionally random-effects procedures (which correspond to fixed-effects procedures when the between-studies variance component is not statistically significant and random-effects procedures when it is significant) generally have performance in between that of fixed- and random-effects procedures. If the inference one wishes to make is not entirely clear, conditionally random-effects procedures might be a reasonable compromise between fixed- and random-effects procedures. However, we believe that clarifying the inference desired (followed by the choice of either fixed- or random-effects procedures) is generally more desirable.

References

Camilli, G. (1990). The test of homogeneity for 2 × 2 contingency tables: A review of and some personal opinions on the controversy. *Psychological Bulletin, 108,* 135–145.

Chang, L. (1992). *A power analysis of the test of homogeneity in effect size meta-analysis.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7,* 177–188.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 106–128.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388–395.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics, 17,* 279–296.

Hedges, L. V., & Olkin, I. (1983). Clustering estimates of effect magnitude from independent studies. *Psychological Bulletin, 93,* 563–573.

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

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hyde, J. S. (1981). How large are cognitive gender differences: A meta-analysis using omega and d. *American Psychologist, 36,* 892–901.

Olkin, I., & Sampson, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics, 54,* 272–277.

Patnaik, P. B. (1949). The noncentral χ^2- and F-distributions and their applications. *Biometrika, 36,* 202–232.

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

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504.

SAS 6.12 [Computer software]. (1996). Cary, NC: The SAS Institute.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Schmidt, F. L., & Hunter, J. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 65,* 643–661.

SPSS 8.0 for Windows [Computer software]. (1998). Chicago: SPSS Inc.

(Appendix follows)

Appendix

Derivations

Bias in the Estimator $\hat{\tau}^2$ of τ^2

When the effect-size parameters are normally distributed and all the conditional variances are equal, that is, $v_1 = \ldots = v_k = v$, then the T_is are independently and identically distributed as $N(\mu, v + \tau^2)$ and the (estimated) weights $w_i^* = w^* = 1/(v + \hat{\tau}^2)$. Consequently, the homogeneity statistic Q is distributed as $[(v + \tau^2)/v]$ multiplied by a central chi-square variate with $(k - 1)$ degrees of freedom (Patnaik, 1949). To obtain the bias in the estimator $\hat{\tau}^2$ given in Equation 10, note that $\hat{\tau}^2$ is a function of Q given by

$$\hat{\tau}^2\,(Q) = \begin{cases} vQ/\,(k-1) - v & \text{if } Q \geqslant k-1 \\ 0 & \text{if } Q < k-1. \end{cases}$$

Using the fact that the probability density of Q is $[v/(v + \tau^2)]g[vx/(v + \tau^2)]$, where $g(x)$ is the probability density function of a chi-square with $k - 1$ degrees of freedom, the expected value of $\hat{\tau}^2$ is

$$E(\hat{\tau}^2) = \frac{v}{v+\tau^2}\int_{k-1}^{\infty}\left(\frac{vq}{k-1} - v\right) g\left(\frac{vq}{v+\tau^2}\right) dq,$$

where $g(x)$ is the probability density function of a chi-square with $k - 1$ degrees of freedom. The values in Table 2 were obtained by numerically integrating the expression above.

Confidence Intervals and Rejection Rates in the Conditional Model

Fixed-Effects Procedures

When all the conditional variances are equal, that is, $v_1 = \ldots = v_k = v$, then the T_is are independently distributed as $N(\theta_i, v)$, and mean $\overline{T}$ is distributed as $N(\overline{\theta}, v/k)$. Consequently, the probability p that $\overline{\theta}$ lies between the confidence limits L and U given in Equation 5 is simply

$$p = P(L < \overline{\theta} < U) = P(\overline{\theta} < U) - P(\overline{\theta} < L),$$

which after calculating the probabilities gives

$$p = 2\Phi(z_{\alpha/2}) - 1 = 1 - \alpha.$$

The rejection rate of the test of the hypothesis that $\overline{\theta} = 0$ is given by $p = P[\overline{T}/\sqrt{(v/k)} > z_\alpha] = 1 - \Phi[z_\alpha - \overline{\theta}\sqrt{(k/v)}]$.

Random-Effects Procedures

When all the conditional variances are equal, that is $v_1 = \ldots = v_k = v$, the (estimated) weights are $w_i^* = w^* = 1/(v + \hat{\tau}^2)$, and consequently $\overline{T}_{\bullet}^*$ is simply the unweighted mean $\overline{T}$ and $v_{\bullet}^* = (v + \hat{\tau}^2)/k$. Using Equation 10 for $\hat{\tau}^2$, the expression for $v_{\bullet}^*$ conditional on Q becomes

$$v_{\bullet}^*\;(Q) = \begin{cases} vQ/k\,(k-1) & \text{if } Q \geqslant k-1 \\ v/k & \text{if } Q < k-1. \end{cases}$$

The sampling distribution of the homogeneity statistic Q is a noncentral chi-square variate with $(k - 1)$ degrees of freedom, and noncentrality parameter

$$\lambda = \sum_{i=1}^{k}(\theta_i - \overline{\theta})^2/v = k\tau^2/v,$$

and is independent of $\overline{T}$.

The $100(1-\alpha)\%$ nominal confidence interval for $\overline{\theta}$ is given by

$$L^* = \overline{T} - z_{\alpha/2}\sqrt{v_{\bullet}^*} < \overline{\theta} < \overline{T} + z_{\alpha/2}\sqrt{v_{\bullet}^*} = U^*.$$

Evaluating the probability content conditional on Q gives

$$P(L^* < \overline{\theta} < U^*|Q) = P(\overline{\theta} < U(Q)|Q) - P(\overline{\theta} < L(Q)|Q).$$

Noting that $\overline{T} \sim N(\overline{\theta}, v/k)$ and evaluating these conditional probabilities gives

$$p(Q) = 2\Phi\{z_{\alpha/2}\sqrt{[kv_{\bullet}^*\;(Q)/v]}\} - 1.$$

Integrating over the distribution of $p(Q)$ and taking account of the functional dependence of $v_{\bullet}^*\;(Q)$ on Q given above yields

$$p = [2\Phi(C_{\alpha/2}) - 1]\int_0^{k-1} f(q|\lambda)\,dq + \int_{k-1}^{\infty}\left[2\Phi\left(C_{\alpha/2}\sqrt{\frac{q}{(k-1)}}\right) - 1\right] f(q|\lambda)\,dq,$$

where $f(q|\lambda)$ is the probability density function of Q, which is a noncentral chi-square with $(k - 1)$ degrees of freedom and noncentrality parameter $\lambda = k\tau^2/v$. The results in Table 3 are obtained by numerically integrating the expression above.

The rejection rate of the test of the hypothesis that $\overline{\theta} = 0$ can also be evaluated conditionally on Q. The conditional one-tailed rejection rate is just

$$P\{Z^* > z_\alpha|Q\} = P\{\sqrt{k}(\overline{T} - \overline{\theta})/\sqrt{v} > \sqrt{k}[z_\alpha\sqrt{v_{\bullet}^*\;(Q)} - \overline{\theta}]/\sqrt{v}\,|Q\}.$$

Evaluating these conditional probabilities using the fact that $\overline{T} \sim N(\overline{\theta}, v/k)$ yields

$$\left[1 - \Phi\left(C_\alpha - \frac{\overline{\theta}\sqrt{k}}{\sqrt{v}}\right)\right]\int_0^{k-1} f(q|\lambda)\,dq + \int_{k-1}^{\infty}\left[1 - \Phi\left(C_\alpha\sqrt{\frac{q}{(k-1)}} - \frac{\overline{\theta}\sqrt{k}}{\sqrt{v}}\right)\right] f(q|\lambda)\,dq,$$

where $f(q|\lambda)$ is the probability density function of Q. The values in Tables 4 and 5 were obtained by numerically integrating the expression above.

Conditionally Random-Effects Procedures

In the conditionally random effects procedures, the random-effects procedures are used if Q is statistically significant, and the fixed-effects procedures, are used otherwise. Thus, the expression for $v.^C$ conditional on Q becomes

$$v.^C(Q)=\begin{cases} vQ/k\,(k-1) & \text{if } Q \geqslant Q_{.95}\\ v/k & \text{if } Q < Q_{.95},\end{cases}$$

where $Q_{.95}$ is the 95th percentile point of the chi-square distribution with $k - 1$ degrees of freedom. The sampling distribution of the homogeneity statistic Q is a noncentral chi-square variate, with $(k - 1)$ degrees of freedom and noncentrality parameter

$$\lambda = \sum_{i=1}^{k} (\theta_i - \overline{\theta})^2/v = k\tau^2/v,$$

and is independent of $\overline{T}$.

The 100(1–α)% nominal confidence interval for $\overline{\theta}$ is given by

$$L^C = \overline{T} - z_{\alpha/2}\sqrt{v.^C} < \overline{\theta} < \overline{T} + z_{\alpha/2}\sqrt{v.^C} = U^C.$$

Evaluating the probability content conditional on Q gives

$$P\{L^C < \overline{\theta} < U^C|Q\} = P\{\overline{\theta} < U(Q)|Q\} - P\{\overline{\theta} < L(Q)|Q\}.$$

Noting that $\overline{T} \sim N\,(\overline{\theta}, v/k)$ and evaluating these conditional probabilities gives

$$p(Q) = 2\Phi\{z_{\alpha/2}\sqrt{[kv.^C\,(Q)/v)]}\} - 1.$$

Integrating over the distribution of $p(Q)$ and taking account the functional dependence of $v.^*\,(Q)$ on Q, given above, yields

$$p = [2\Phi(C_{\alpha/2}) - 1]\int_0^{Q.95} f(q|\lambda)\,dq + \int_{Q.95}^{\infty}\left[2\Phi\left(C_{\alpha/2}\sqrt{\frac{q}{(k-1)}}\right) - 1\right] f(q|\lambda)\,dq,$$

where $f(q|\lambda)$ is the probability density function of Q (which is just a noncentral chi-square with $(k - 1)$ degrees of freedom and noncentrality parameter $\lambda = k\tau^2/v$). The results in Table 3 are obtained by numerically integrating the expression above.

The rejection rate of the test of the hypothesis that $\overline{\theta} = 0$ can also be evaluated conditionally on Q. The conditional one-tailed rejection rate is just

$$\begin{aligned} &P\{Z^C > z_\alpha\,|Q\}\\ &\quad = P\{\sqrt{k}(\overline{T} - \overline{\theta})/\sqrt{v.^C\,(Q)}\\ &\quad > \sqrt{k}[z_\alpha\sqrt{v.^C\,(Q)} - \overline{\theta}]/\sqrt{v}\,|Q\}.\end{aligned}$$

Evaluating these conditional probabilities using the fact that $\overline{T} \sim N[\overline{\theta}, v/k]$ yields

$$\left[1 - \Phi\left(C_\alpha - \frac{\overline{\theta}\sqrt{k}}{\sqrt{v}}\right)\right]\int_0^{Q.95} f(q|\lambda)\,dq + \int_{Q.95}^{\infty}\left[1 - \Phi\left(C_\alpha\sqrt{\frac{q}{(k-1)}} - \frac{\overline{\theta}\sqrt{k}}{\sqrt{v}}\right)\right] f(q|\lambda)\,dq,$$

where $f(q|\lambda)$ is the probability density function of Q. The values in Tables 4 and 5 were obtained by numerically integrating the expression above.

Confidence Intervals and Rejection Rates in the Unconditional Model

Fixed-Effects Procedures

When the effect-size parameters are normally distributed and all the conditional variances are equal, that is $v_1 = \ldots = v_k = v$, then the T_is are independently distributed as $N(\mu, v + \tau^2)$, and mean $\overline{T}$ is distributed as $N[\mu, (v + \tau^2)/k]$. Consequently, the probability p that μ lies between the confidence limits L and U given in Equation 5 is simply

$$p = P(L < \mu < U) = P(\mu < U) - P(\mu < L),$$

which after evaluating the probabilities gives

$$p = 2\Phi\{z_{\alpha/2}\sqrt{[v/(v + \tau^2)]}\} - 1.$$

The rejection rate of the test of the hypothesis that $\mu = 0$ is given by

$$\begin{aligned} p &= P\{\overline{T}/\sqrt{[v/k]} > z_\alpha\}\\ &= 1 - \Phi\{z_\alpha\sqrt{[v/(v + \tau^2)]} - \mu\sqrt{[k/(v + \tau^2)]}\}.\end{aligned}$$

Random-Effects Procedures

When all the conditional variances are equal, that is $v_1 = \ldots = v_k = v$, the (estimated) weights are $w_i^* = w^* = 1/(v + \hat{\tau}^2)$, and consequently $\overline{T}.^*$ is simply the unweighted mean $\overline{T}.$, and $v.^* = (v + \hat{\tau}^2)/k$. Using Equation 10 for $\hat{\tau}^2$, the expression for $v.^*$ conditional on Q becomes

$$v.^*\,(Q)=\begin{cases} vQ/k\,(k-1) & \text{if } Q \geqslant k-1\\ v/k & \text{if } Q < k-1.\end{cases}$$

The sampling distribution of the homogeneity statistic Q is $(v + \tau^2)/v$ multiplied by a central chi-square variate, with $(k - 1)$ degrees of freedom, and is independent of $\overline{T}$.

The 100(1–α)% nominal confidence interval for μ is given by

$$L^* = \overline{T} - z_{\alpha/2}\sqrt{v.^*} < \overline{T} + z_{\alpha/2}\sqrt{v.^*} = U^*.$$

Evaluating the probability content conditional on Q gives

$$P[L^* < \mu < U^*|Q] = P[\mu < U(Q)|Q] - P[\mu < L(Q)|Q].$$

Noting that $\overline{T} \sim N(\mu, (v + \tau^2)/k)$ and evaluating these conditional probabilities give

$$p(Q) = 2\Phi\{z_{\alpha/2}\surd[kv.^* (Q)]/\surd(v + \tau^2)\} - 1.$$

Integrating over the distribution of $p(Q)$ and taking account of the functional dependence of $v.^* (Q)$ on Q given above yield

$$p = \left[2\Phi\left(C_{\alpha/2}\sqrt{\frac{v}{v+\tau^2}}\right) - 1\right]\int_0^{k-1} f(q)\,dq + \int_{k-1}^{\infty}\left[2\Phi\left(C_{\alpha/2}\sqrt{\frac{vq}{(k-1)(v+\tau^2)}}\right) - 1\right] f(q)\,dq,$$

where $f(q)$ is the probability density function of Q. The results in Table 6 are obtained by numerically integrating the expression above.

The rejection rate of the test of the hypothesis that $\mu = 0$ can also be evaluated conditionally on Q. The conditional one-tailed rejection rate is just

$$\begin{aligned}&P\{Z^* > z_\alpha|Q\}\\ &\quad = P\{\surd k(\overline{T} - \mu)/\surd(v + \tau^2)\\ &\quad > \surd k[z_\alpha \surd v.^* (Q) - \mu]/\surd(v + \tau^2)\,|Q\}.\end{aligned}$$

Evaluating these conditional probabilities using the fact that $\overline{T} \sim N[\overline{\theta}, (v + \tau^2)/k]$ yields

$$\left[1 - \Phi\left(C_\alpha\sqrt{\frac{v}{v+\tau^2}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right]\int_0^{k-1} f(q)\,dq + \int_{k-1}^{\infty}\left[1 - \Phi\left(C_\alpha\sqrt{\frac{vq}{(v+\tau^2)(k-1)}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] f(q)\,dq,$$

where $f(q)$ is the probability density function of Q. The values in Tables 7 and 8 were obtained by numerically integrating the expression above.

Conditionally Random-Effects Procedures

In the conditionally random effects procedure, the random-effects procedures are used if Q is statistically significant, and the fixed-effects procedures are used otherwise. Thus, the expression $v.^C$ for the variance of $\overline{T}$ conditional on Q becomes

$$v.^C(Q) = \begin{cases} vQ/k\,(k-1) & \text{if } Q \geqslant Q_{.95} \\ v/k & \text{if } Q < Q_{.95}, \end{cases}$$

where Q_α is the $100\alpha\%$ point of the chi-square distribution with $(k - 1)$ degrees of freedom. The sampling distribution of the homogeneity statistic Q is $(v + \tau^2)/v \times$ a central chi-square variate, with $(k - 1)$ degrees of freedom, and is independent of $\overline{T}$.

The $100(1-\alpha)\%$ nominal confidence interval for μ is given by

$$L^C = \overline{T} - z_{\alpha/2}\surd v.^C < \mu < \overline{T} + z_{\alpha/2}\surd v.^C = U^C.$$

Evaluating the probability content conditional on Q gives

$$P[L^C < \mu < U^C|Q] = P[\mu < U(Q)|Q] - P[\mu < L(Q)|Q].$$

Noting that $\overline{T} \sim N(\mu, (v + \tau^2)/k)$ and evaluating these conditional probabilities give

$$p(Q) = 2\Phi\{z_{\alpha/2}\surd[kv.^C(Q)]/\surd(v + \tau^2)\} - 1.$$

Integrating over the distribution of $p(Q)$ and taking account of the functional dependence of $v.^C (Q)$ on Q given above yield

$$p = \left[2\Phi\left(C_{\alpha/2}\sqrt{\frac{v}{v+\tau^2}}\right) - 1\right]\int_0^{Q.95} f(q)\,dq + \int_{Q.95}^{\infty}\left[2\Phi\left(C_{\alpha/2}\sqrt{\frac{vq}{(k-1)(v+\tau^2)}}\right) - 1\right] f(q)\,dq,$$

where $f(q)$ is the probability density function of Q and Q_α is defined above. The results in Table 6 are obtained by numerically integrating the expression above.

The rejection rate of the test of the hypothesis that $\mu = 0$ can also be evaluated conditionally on Q. The conditional one-tailed rejection rate is just

$$\begin{aligned}&P\{Z^* > z_\alpha\,|Q\}\\ &\quad = P\{\surd k(\overline{T} - \mu)/\surd(v + \tau^2)\\ &\quad > \surd k[z_\alpha \surd v.^* (Q) - \mu]/\surd(v + \tau^2)\,|Q\}.\end{aligned}$$

Evaluating these conditional probabilities using the fact that $\overline{T} \sim N[\overline{\theta}, (v + \tau^2)/k]$ yields

$$\left[1 - \Phi\left(C_\alpha\sqrt{\frac{v}{v+\tau^2}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right]\int_0^{Q.95} f(q)\,dq + \int_{Q.95}^{\infty}\left[1 - \Phi\left(C_\alpha\sqrt{\frac{vq}{(v+\tau^2)(k-1)}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] f(q)\,dq,$$

where $f(q)$ is the probability density function of Q. The values in Tables 7 and 8 were obtained by numerically integrating the expression above.

Received December 19, 1997
Revision received March 16, 1998
Accepted July 5, 1998 ■