DISCUSSION PAPER SERIES

IZA DP No. 2198

Fixed and Mixed Effects Models in Meta-Analysis

Spyros Konstantopoulos

July 2006

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Fixed and Mixed Effects Models in Meta-Analysis

Spyros Konstantopoulos
Northwestern University
and IZA Bonn

Discussion Paper No. 2198
July 2006

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 2198
July 2006

ABSTRACT

Fixed and Mixed Effects Models in Meta-Analysis

The last three decades the accumulation of quantitative research evidence has led to the development of systematic methods for combining information across samples of related studies. Although a few methods have been described for accumulating research evidence over time, meta-analysis is widely considered as the most appropriate statistical method for combining evidence across studies. This study reviews fixed and mixed effects models for univariate and multivariate meta-analysis. In addition, the study discusses specialized software that facilitates the statistical analysis of meta-analytic data.

JEL Classification: C02

Keywords: meta-analysis, mixed models, multivariate analysis

Corresponding author:

Spyros Konstantopoulos
School of Education and Social Policy
Northwestern University
2120 Campus Drive
Evanston, IL 60208
USA
E-mail: spyros@northwestern.edu

The last three decades the growth of the social science research enterprise has led to a large body of related research studies, which poses the question of how to organize and summarize these findings in order to identify and exploit what is known, and focus research on promising areas. This accumulation of quantitative research evidence has led to the development of systematic methods for combining information across samples of related studies. Although a few methods have been described for accumulating research evidence over time, meta-analysis (e.g., Cooper, & Hedges, 1994; Hedges & Olkin, 1985) is widely considered the most popular and most appropriate.

Meta-analysis has recently gained ample recognition in the statistical, medical, and social science communities. Meta-analysis refers to quantitative methods of synthesizing empirical research evidence from a sample of studies that examine a certain topic and test comparable hypotheses (Hedges & Olkin, 1985). The first step in meta-analysis involves describing the results of each study via numerical indicators (e.g., estimates of effect sizes such as a standardized mean difference, a correlation coefficient, or an odds ratio). These effect size estimates reflect the magnitude of the association of interest in each study. The second step involves combining the effect size estimates from each study to produce a single indicator that summarizes the relationship of interest across the sample of studies. Hence, meta-analytic procedures produce summary statistics, which are then tested to determine their statistical significance and importance.

The specific analytic techniques involved will depend on the question the meta-analytic summary is intended to address. Sometimes the question of interest concerns the typical or average study result. For example in studies that measure the effect of some treatment or intervention, the average effect of the treatment is often of interest (see, e.g., Smith and Glass, 1977). In other cases the degree of variation in

results across studies will be of primary interest. For example, meta-analysis is often used to study the generalizability of employment test validities across situations (see, e.g., Schmidt and Hunter, 1977). In other cases, the primary interest is in the factors that are related to study results. For example, meta-analysis is often used to identify the contexts in which a treatment or intervention is most successful or has the largest effect (see, e.g., Cooper, 1989).

One advantage of employing meta-analysis is that the pooled results generated can verify or refute theories, and therefore can facilitate the improvement of substantive theory. In addition, from a statistical point of view the results of meta-analytic procedures have higher statistical power than indicators obtained from individual studies, which increases the probability of detecting the associations of interest (Cohn & Becker, 2003). A substantial advantage, however, of meta-analysis is the generality of results across studies. This constitutes a unique aspect of research synthesis that is crucial for the external validation of the estimates (see Shadish, Cook, & Campbell, 2002). Generally, the estimates that are produced from meta-analyses higher external validity than estimates reported in single studies.

The term meta-analysis is sometimes used to describe the entire process of quantitative research synthesis. However, more recently, it has been used specifically for the statistical component of research synthesis. In this study we deal exclusively with the narrower usage of the term to describe statistical methods only. Nonetheless, it is crucial to understand that in research synthesis, as in any research, statistical methods are only one part of the enterprise. Statistical methods cannot remedy the problem of data, which are of poor quality. Excellent treatments of the non-statistical aspects of research synthesis are available in Cooper (1989), Cooper and Hedges (1994), and (Lipsey and Wilson, 2001).

Effect Size Estimates

Effect sizes are quantitative indexes that are used to summarize the results of a study in meta-analysis. That is, effect sizes reflect the magnitude of the association between variables of interest in each study. There are many different effect sizes and the effect size used in a meta-analysis should be chosen so that it represents the results of a study in a way that is easily interpretable and is comparable across studies. In a sense, effect sizes should put the results of all studies "on a common scale" so that they can be readily interpreted, compared, and combined. It is important to distinguish the effect size estimate in a study from the effect size parameter (the true effect size) in that study. In principle, the effect size estimate will vary somewhat from sample to sample that might be obtained in a particular study. The effect size parameter is in principle fixed. One might think of the effect size parameter as the estimate that would be obtained if the study had a very large (essentially infinite) sample, so that the sampling variation is negligible.

The choice of an effect size index will depend on the design of the studies, the way in which the outcome is measured, and the statistical analysis used in each study. Most of the effect size indexes used in the social sciences will fall into one of three families of effect sizes: the standardized mean difference family, the odds ratio family, and the correlation coefficient family.

The standardized mean difference

In many studies of the effects of a treatment or intervention that measure the outcome on a continuous scale, a natural effect size is the standardized mean difference. The standardized mean difference is the difference between the mean outcome in the treatment group and the mean outcome in the control group divided by the within group standard deviation. That is the standardized mean difference is

$$d = \frac{\bar{Y}^T - Y^C}{S},$$

where $\bar{Y}^T$ is the sample mean of the outcome in the treatment group, $\bar{Y}^C$ is the sample mean of the outcome in the control group, and S is the within-group standard deviation of the outcome. The corresponding standardized mean difference parameter is

$$\delta = \frac{\mu^T - \mu^C}{\sigma},$$

where μ^T is the population mean in the treatment group, μ^C is the population mean outcome in the control group, and σ is the population within-group standard deviation of the outcome. This effect size is easy to interpret since it is just the treatment effect in standard deviation units. It can also be interpreted as having the same meaning across studies (see Hedges and Olkin, 1985). The sampling uncertainty of the standardized mean difference is characterized by its variance which is

$$v = \frac{n^T + n^C}{n^T n^C} + \frac{d^2}{2(n^T + n^C)},$$

where n^T and n^C are the treatment and control group sample sizes, respectively. Note that this variance can be computed from a single observation of the effect size if the sample sizes of the two groups within a study are known. Because the standardized mean difference is approximately normally distributed, the square root of the variance (the standard error) can be used to compute confidence intervals for the true effect size or effect size parameter δ. Specifically, a 95% confidence interval for the effect size is given by

$$d - 2\sqrt{v} \leq \delta \leq d + 2\sqrt{v}.$$

Several variations of the standardized mean difference are also sometimes used as effect sizes (see Rosenthal, 1994).

The log odds ratio

In many studies of the effects of a treatment or intervention that measure the outcome on a

dichotomous scale, a natural effect size is the log odds ratio. The log odds ratio is just the log of the ratio of the odds of a particular one of the two outcomes (the target outcome) in the treatment group to the odds of that particular outcome in the control group. That is, the log odds ratio is

$$\log(\text{OR}) = \log\left(\frac{p^T/(1-p^T)}{p^C/(1-p^C)}\right) = \log\left(\frac{p^T(1-p^C)}{p^C(1-p^T)}\right),$$

where p^T and p^C are the proportion of the treatment and control groups, respectively that have the target outcome. The corresponding odds ratio parameter is

$$\omega = \log\left(\frac{\pi^T/(1-\pi^T)}{\pi^C/(1-\pi^C)}\right) = \log\left(\frac{\pi^T(1-\pi^C)}{\pi^C(1-\pi^T)}\right),$$

where π^T and π^C are the population proportions in the treatment and control groups, respectively, that have the target outcome. The log odds ratio is widely used in the analysis of data that have dichotomous outcomes and is readily interpretable by researchers who frequently encounter this kind of data. It also has the same meaning across studies so it is suitable for combining (see Fleiss, 1994).

The sampling uncertainty of the log odds ratio is characterized by its variance, which is

$$v = \frac{1}{n^T p^T} + \frac{1}{n^T(1-p^T)} + \frac{1}{n^C p^C} + \frac{1}{n^C(1-p^C)},$$

where n^T and n^C are the treatment and control group sample sizes, respectively. As in the case of the standardized mean difference, the log odds ratio is approximately normally distributed, and the square root of the variance (the standard error) can be used to compute confidence intervals for the true effect size or effect size parameter ω. Specifically, a 95% confidence interval for the effect size is given by

$$d - 2\sqrt{v} \leq \omega \leq d + 2\sqrt{v}\ .$$

There are several other indexes in the odds ratio family, including the risk ratio (the ratio of

proportion having the target outcome in the treatment group to that in the control group or p^T/p^C) and the risk difference (the difference between the proportion having a particular one of the two outcomes in the treatment group and that in the control group or $p^T - p^C$). For a discussion of effect size measures for studies with dichotomous outcomes, including the the odds ratio family of effect sizes, see Fleiss (1994).

The correlation coefficient

In many studies of the relation between two continuous variables, the correlation coefficient is a natural measure of effect size. Often this correlation is transformed via the Fisher z-transform

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

in carrying out statistical analyses. The corresponding correlation parameter is ρ, the population correlation and the parameter that corresponds to the estimate z is ζ, the z-transform of ρ. The sampling uncertainty of the z-transformed correlation is characterized by its variance

$$v = \frac{1}{n-3},$$

where n is the sample size of the study, and it is used in the same way as are the variances of the standardized mean difference and log odds ratio to obtain confidence intervals.

The statistical methods for meta-analysis are quite similar, regardless of the effect size measure used. Therefore, in the rest of this chapter we do not describe statistical methods that are specific to a particular effect size index, but describe them in terms of a generic effect size measure T_i. We assume that the T_i are normally distributed about the corresponding θ_i with known variance v_i. That is, we assume that

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

This assumption is very nearly true for effect sizes such as the Fisher z-transformed correlation coefficient

and standardized mean differences. However for effect sizes such as the untransformed correlation coefficient, or the log-odds ratio, the results are not exact, but remain true as large sample approximations. For a discussion of effect size measures for studies with continuous outcomes, see Rosenthal (1994) and for a treatment of effect size measures for studies with categorical outcomes see Fleiss (1994).

Fixed Effects Models

Two somewhat different statistical models have been developed for inference about effect size data from a collection of studies, called the fixed effects and the mixed (or random) effects models (see, e.g., Hedges and Vevea, 1998). 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 size parameters (see e.g., Hedges, 1982, 1994; Rosenthal & Rubin, 1982). The logic of fixed effects models is that inferences are not about any hypothesized population of studies, but about the particular collection of studies that is observed.

The simplest fixed effects model involves the estimation of an average effect size by combining the effect size estimates across all studies in the sample. Let θ_i be the (unobserved) effect size parameter (the true effect size) in the i^{th} study, let T_i be the corresponding observed effect size estimate from the i^{th} study, and let v_i be its variance. Thus the data from a set of k studies are the effect size estimates $T_1, \ldots, T_k$ and their variances $v_1, \ldots, v_k$.

The effect size estimate T_i is modeled as the effect size parameter plus a sampling error ε_i. That is

$$T_i = \theta_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, v_i).$$

The parameter θ is the mean effect size parameter for all of the studies. It has the interpretation that θ is the mean of the distribution from which the study-specific effect size parameters ($\theta_1, \theta_2, \ldots, \theta_k$) were

sampled. Note that this is not conceptually the same as the mean of $\theta_1, \theta_2, ..., \theta_k$, the effect size parameters of the k studies that were observed. The effect size parameters are in turn determined by a mean effect size β_0, that is

$$\theta_i = \beta_0,$$

which indicates that the θ_i's are fixed and thus

$$T_i = \beta_0 + \varepsilon_i. \qquad (1)$$

Note that in meta-analysis, the variances (the v_i's) are different for each of the studies. That is, each study has a *different* sampling error variance. In addition, in meta-analysis these variances are known. Since, the amount of sampling uncertainty is not identical in every study, it seems reasonable that, if an average effect size is to be computed across studies, it would be desirable to give more weight in that average to studies that have more precise estimates (or smaller variances) than those with less precise estimates.

The weighted least squares (and maximum likelihood) estimate of β_0 under the model is

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \qquad (2)$$

where $w_i = 1/v_i = 1/v_i$. Note that this estimator, corresponds to a weighted mean of the T_i, giving more weight to the studies whose estimates have smaller unconditional variance (are more precise) when pooling. This is actually a weighted regression including only the constant term (intercept).

The sampling variance $v_\bullet$ of $\hat{\beta}_0$ is simply the reciprocal of the sum of the weights,

$$v_\bullet = \left(\sum_{i=1}^{k} w_i \right)^{-1},$$

and the standard error SE($\hat{\beta}_0$) of $\hat{\beta}_0$ is just the square root of v.. Under this model $\hat{\beta}_0$ is normally distributed so a 100(1-α) percent confidence interval for β_0 is given by

$$\hat{\beta}_0 - t_{\alpha/2}\sqrt{v_\bullet} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2}\sqrt{v_\bullet},$$

where t_α is the 100α percent point of the t-distribution with (k - 1) degrees of freedom. Similarly, a two-sided test of the hypothesis that $\beta_0 = 0$ at significance level α uses the test statistic $Z = \hat{\beta}_0 / \sqrt{v_\bullet}$ and rejects if |Z| exceeds $t_{\alpha/2}$. Note that the same test and confidence intervals can be computed for any individual coefficient (when multiple predictors are included in the regression model in equation 1).

A more general fixed effects model includes predictors in the regression equation. Suppose that there are k studies and that in each study there are p predictors. Then the effect size parameter θ_i for the i^{th} study depends on $\mathbf{x}_i = (x_{i1},...,x_{ip})'$ via a linear model

$$\theta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \; i = 1, ..., k, \qquad (3)$$

where $\beta_1,...,\beta_p$ are unknown regression coefficients. Across all studies the k x p matrix **X**

$$\mathbf{X} = \begin{bmatrix} x_{11}\, x_{12} \ldots x_{1p} \\ x_{21}\, x_{22} \ldots x_{2p} \\ . \;\; . \;\; \ldots \;\; . \\ . \;\; . \;\; \ldots \;\; . \\ x_{k1}\, x_{k2} \ldots x_{kp} \end{bmatrix}$$

is called the *design matrix* which is assumed to have no linearly dependent columns; that is, **X** has rank p. It is often convenient to define $x_{11} = x_{21} = ... = x_{k1} = 1$, so that the first regression coefficient becomes a

constant term (intercept), as in ordinary regression.

The model in equation (3) can be written in matrix notation as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ is the p dimensional vector of regression coefficients.

Thus, the model for **T** can then be written in a typical regression form as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (4)$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_k)' = \mathbf{T} - \boldsymbol{\theta}$ is a k dimensional vector of residuals. Because $\boldsymbol{\varepsilon} = \mathbf{T} - \boldsymbol{\theta}$, it follows that the distribution of $\boldsymbol{\varepsilon}$ is approximately a k-variate normal distribution with means zero and diagonal covariance matrix **V** given by

$$\mathbf{V} = \text{Diag}(v_1, v_2, ..., v_k),$$

which indicates that the elements of $\boldsymbol{\varepsilon}$ are independent but not identically distributed. Therefore we can use the method of generalized least squares to obtain an estimate of $\boldsymbol{\beta}$.

The generalized least squares estimator $\widehat{\boldsymbol{\beta}}$ under the model in 4, which is also the maximum likelihood estimator of $\boldsymbol{\beta}$ under that model, is given by

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{T} \qquad (5)$$

which has a normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}$ given by

$$Cov(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}. \qquad (6)$$

In cases where the effect size estimates are not exactly normally distributed or the variance v_i is not known exactly (e. g., when it depends on the unknown effect size), the matrix **V** depends on the unknown parameter $\boldsymbol{\theta}$. However, it is still possible to estimate $\boldsymbol{\beta}$ by substituting the estimated

variance $\hat{v}_i$ for v_i in **V** and

using the estimated covariance matrix to carry out the generalized least squares estimate. The resulting estimator is not the maximum likelihood estimator, but has the same asymptotic distribution as the maximum likelihood estimator of **β** as the total sample size tends to infinity at the same rate in all studies (see Hedges, 1983).

Tests for Blocks of Regression Coefficients

In the fixed effects model, a researcher sometimes want to test whether a subset $\beta_1, ..., \beta_m$ of the regression coefficients are simultaneously zero, that is,

H_0: $\beta_1 = ... = \beta_m = 0$.

This test arises, for example in stepwise analyses where it is desired to determine whether a set of m of the p predictor variables ($m \leq p$) are related for effect size after controlling for the effects of the other predictor variables. For example, suppose one is interested in testing the importance of a conceptual variable such as research design, which is coded as a set of predictors. Specifically, such a variable can be coded as multiple dummies for randomized experiment, matched samples, non-equivalent comparison group samples, and other quasi-experimental designs, but it's treated as one conceptual variable and its importance is tested simultaneously. To test this hypothesis, we compute the statistic

$$Q = (\hat{\beta}_1, \ldots, \hat{\beta}_m)(\mathbf{\Sigma_{11}})^{-1}(\hat{\beta}_1, \ldots, \hat{\beta}_m)' \qquad (7)$$

where $\mathbf{\Sigma_{11}}$ is the variance-covariance matrix of the m regression coefficients. The test that $\beta_1 = ... = \beta_m = 0$ at the 100α-percent significance level consists in rejecting the null hypothesis if Q exceeds the $100(1 - \alpha)$ percentage point of the chi-square distribution with m degrees of freedom. If $m = p$,

then the procedure above yields a test that all the β_j are simultaneously zero, that is $\boldsymbol{\beta} = \mathbf{0}$. In this case the test statistic Q given in (7) becomes the weighted sum of squares due to regression

$$Q_R = \hat{\boldsymbol{\beta}}' \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}.$$

The test that $\boldsymbol{\beta} = \mathbf{0}$ is simply a test of whether the weighted sum of squares due to the regression is larger than would be expected if $\boldsymbol{\beta} = \mathbf{0}$, and the test consists of rejecting the hypothesis that $\boldsymbol{\beta} = \mathbf{0}$ if Q_R exceeds the $100(1 - \alpha)$ percentage point of a chi-square with p degrees of freedom.

Example

Gender differences in field articulation ability (sometimes called visual-analytic spatial ability) were studied by Hyde (1981). She reported standardized mean differences from 14 studies that examined gender differences in spatial ability tasks that call for the joint application of visual and analytic processes (see Maccoby & Jacklin, 1974). All estimates are positive and indicate that on average males performed higher than females in field articulation. The effect size estimates are reported in column two of Table 1. The variances of the effect size estimates are reported in column three. The year the study was conducted is in column four.

First, we compute the weighted mean of the effect size estimates. This yields an overall mean estimate of $\hat{\beta}_0 = 0.547$ with a variance of $v_\bullet = 0.0046$. The 95% confidence interval for β_0 is given by $0.4005 = 0.547 - 2.160\sqrt{0.0046} \leq \beta_0 \leq 0.547 - 2.160\sqrt{0.0046} = 0.6935$. This confidence interval does not include zero, so the data are incompatible with the hypothesis that $\beta_0 = 0$. Alternatively, the ratio $\hat{\beta}_0 / v_\bullet = 8.05$ which indicates that the overall mean is significantly different from zero since the observed value is larger than the two-tailed critical t-value at the 0.05 significance level with 13

degrees of freedom (2.160).

Second we compute the effect of the year of study. This yields an estimate of $\widehat{\beta}_1$ = -0.0433 with a variance of $\text{var}\left(\widehat{\beta}_1\right) = 0.0002$. The 95% confidence interval for β_1 is given by -0.0741 = -0.0433 - 2.179$\sqrt{0.0002} \leq \beta_1 \leq$ -0.0433 + 2.179$\sqrt{0.0002}$ = -0.0125. This confidence interval does not include 0, so the data are incompatible with the hypothesis that $\beta_1 = 0$. Alternatively, the ratio $\widehat{\beta}_1 / \text{var}\left(\widehat{\beta}_1\right) = -2.999$ which indicates that the year of the study effect is significantly different from zero since the absolute observed value is larger than the two-tailed critical t-value at the 0.05 significance level with 12 degrees of freedom (2.179). This indicates that the effect size estimates get smaller over time. The above results are easily obtained from the second version of comprehensive meta-analysis developed by Hedges et al., (2005)[1].

Mixed Effects Models

Mixed effects models treat the effect size parameters as if they were a random sample from a population of effect parameters and estimate hyper-parameters (usually the mean and variance) describing this population of effect parameters (see, e.g., Schmidt and Hunter, 1977; Hedges, 1983; DerSimonian and Laird, 1986). The term mixed effects model is appropriate since the parameter structure of these models is identical to those of the general linear mixed model (and their important application in social sciences, hierarchical linear models).

In this case, there is non-negligible variation among effect size parameters even after controlling for the factors that are of interest in the analysis. That is, there is greater residual variation than would be expected from sampling error alone after controlling for all of the study level covariates. If the researcher believes that this variation should be included in computations of

the uncertainty of the regression coefficient estimates, fixed effects models are *not* appropriate because such excess residual variation has no effect on computation of estimates or their uncertainty in fixed effects models. The mixed effects model is a generalization of the fixed effects model that incorporates a component of between-study variation into the uncertainty of the effect size parameters and their estimates.

As in fixed effects models the simplest mixed effects model involves the estimation of an average effect size by combining the effect size estimates across all studies in the sample. However, in this case a natural way to describe the data is via a two-level model with one model for the data at the study level and another model for the between-study variation in effects. The within-study level is as defined for fixed effects models. In the between-study level, the effect size parameters are determined by a mean effect size β_0 plus a study-specific random effect η_i. That is

$$\theta_i = \beta_0 + \eta_i \qquad \eta_i \sim N(0, \tau^2).$$

In this model, the η_i represent differences between the effect size parameters from study to study. The parameter τ^2, often called the between-study variance component, describes the amount of variation across studies in the random effects (the η_i's) and therefore effect parameters (the θ_i's).

The two-level model described above can be written as a one-level model as follows

$$T_i = \beta_0 + \eta_i + \varepsilon_i = \beta_0 + \xi_i ,$$

where ξ_i is a composite error defined by $\xi_i = \eta_i + \varepsilon_i$. Writing this as a one-level model, we see that each effect size is an estimate of β_0 with a variance that depends on both v_i and τ^2. Hence, 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 the θ_i as well. The latter is the *unconditional sampling variance* of T_i

(denoted v_i^*). Since the sampling error ε_i and the random effect η_i are assumed to be independent and the variance of η_i is $\hat{\tau}^2$, it follows that the unconditional sampling variance of T_i is $v_i^* = v_i + \hat{\tau}^2$.

The least squares (and maximum likelihood) estimate of the mean β_0 under the model is

$$\hat{\beta}_0^* = \frac{\sum_{i=1}^{k} w_i^* T_i}{\sum_{i=1}^{k} w_i^*} \tag{8}$$

where $w_i^* = 1/(v_i + \hat{\tau}^2) = 1/v_i^*$ and $\hat{\tau}^2$ is the between-study variance component estimate. Note that this estimator, corresponds to a weighted mean of the T_i, giving more weight to the studies whose estimates have smaller variance (are more precise) when pooling.

The sampling variance $v_\bullet^*$ of $\hat{\beta}_0^*$ is simply the reciprocal of the sum of the weights,

$$v_\bullet^* = \left(\sum_{i=1}^{k} w_i^* \right)^{-1},$$

and the standard error $SE(\hat{\beta}_0^*)$ of $\hat{\beta}_0^*$ is just the square root of $v_\bullet^*$. Under this model $\hat{\beta}_0^*$ is normally distributed so a 100(1-α) percent confidence interval for β_0 is given by

$$\hat{\beta}_0^* - t_{\alpha/2}\sqrt{v_\bullet^*} \le \beta_0 \le \hat{\beta}_0^* + t_{\alpha/2}\sqrt{v_\bullet^*},$$

where t_α is the 100α percent point of the t-distribution with (k - 1) degrees of freedom. Similarly, a two-sided test of the hypothesis that $\beta_0 = 0$ at significance level α uses the test statistic $Z = \hat{\beta}_0^* / \sqrt{v_\bullet^*}$ and rejects if $|Z|$ exceeds $t_{\alpha/2}$. Note that the same test and confidence intervals can be computed for any individual coefficient (when multiple predictors are included in the regression).

A more general mixed effects model includes predictors in the regression equation. Suppose

that there are k studies and that in each study there are p predictors. Then the effect size parameter θ_i for the i^{th} study depends on $\mathbf{x}_i = (x_{i1},...,x_{ip})'$ via a linear model

$$\theta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \eta_i, \qquad \eta_i \sim N(0, \tau^2)$$

where $x_{i1}, \ldots, x_{ip}$ are the values of the predictor variables $X_1, \ldots, X_p$ for the i^{th} study (that is x_{ij} is the value of predictor variable X_j for study i), η_i is a study-specific random effect with zero expectation and variance τ^2.

Then, the single equation as a model for the T_i is

$$T_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \eta_i + \varepsilon_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \xi_i ,$$

where $\xi_i = \eta_i + \varepsilon_i$ is a composite residual incorporating both study-specific random effect and sampling error. Because we assume that η_i and ε_i are independent, it follows that the variance of ξ_i is $\tau^2 + v_i$. If τ^2 were known, we could estimate the regression coefficients via weighted least squares (which would also yield the maximum likelihood estimates of the β_i's). The description of the weighted least squares estimation is facilitated by describing the model in matrix notation.

We denote the k-dimensional vectors of population and sample effect sizes by $\boldsymbol{\theta} = (\theta_1,...,\theta_k)'$ and $\mathbf{T} = (T_1,...,T_k)'$, respectively. The model for the observations $\mathbf{T}$ as a one level model can be written as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \qquad (9)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1..., \beta_p)'$ is the p-dimensional vector of regression coefficients $\boldsymbol{\eta} = (\eta_1, ..., \eta_k)'$ is the k-dimensional vector of random effects, and $\boldsymbol{\xi} = (\xi_1, ..., \xi_k)'$ is a k-dimensional vector of residuals of $\mathbf{T}$ about $\mathbf{X}\boldsymbol{\beta}$. The covariance matrix of $\boldsymbol{\xi}$ is a diagonal matrix where the i^{th} diagonal element is $v_i + \hat{\tau}^2$.

If the residual variance component τ^2 were known, we could use the method of generalized

least squares to obtain an estimate of **β**. Although we do not know the residual variance component τ^2, we can compute an estimate of τ^2 and use this estimate to obtain a generalized least squares estimate of **β**. The unconditional covariance matrix of the estimates is a k x k diagonal matrix **V*** be defined by

$$\mathbf{V}^* = \text{Diag}(v_1 + \hat{\tau}^2, v_2 + \hat{\tau}^2, ..., v_k + \hat{\tau}^2).$$

The generalized least squares estimator $\hat{\boldsymbol{\beta}}^*$ under the model (9) using the estimated covariance matrix $\hat{\mathbf{V}}^*$ is given by

$$\hat{\boldsymbol{\beta}}^* = \left[\mathbf{X}'(\mathbf{V}^*)^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'(\mathbf{V}^*)^{-1}\mathbf{T} \qquad (10)$$

which, is normally distributed with mean **β** and covariance matrix **Σ*** given by

$$\boldsymbol{\Sigma}^* = \left[\mathbf{X}'(\mathbf{V}^*)^{-1}\mathbf{X}\right]^{-1}. \qquad (11)$$

As equation 11 indicates the estimate of the between study variance component $\hat{\tau}^2$ is incorporated as a constant term in the computation of the regression coefficients and their dispersion via the variance covariance matrix of the effect size estimates.

Tests for Blocks of Regression Coefficients

As in the fixed effects model, we sometimes want to test whether a subset $\beta_1^*, ..., \beta_m^*$ of the regression coefficients are simultaneously zero, that is,

$$H_0: \beta_1^* = ... = \beta_m^* = 0.$$

This test arises, for example in stepwise analyses where it is desired to determine whether a set of m of the p predictor variables ($m \leq p$) are related for effect size after controlling for the effects of the

other predictor variables. To test this hypothesis, compute $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_m^*, \hat{\beta}_{m+1}^*, \ldots, \hat{\beta}_p^*)'$ and the statistic

$$Q^* = (\hat{\beta}_1^*, \ldots, \hat{\beta}_m^*)(\boldsymbol{\Sigma}_{\mathbf{11}}^*)^{-1}(\hat{\beta}_1^*, \ldots, \hat{\beta}_m^*)' \qquad (12)$$

where $\boldsymbol{\Sigma}_{\mathbf{11}}^*$ is the variance-covariance matrix of the m regression coefficients. The test that $\beta_1^* = \ldots = \beta_m^* = 0$ at the 100α-percent significance level consists in rejecting the null hypothesis if Q^* exceeds the 100(1 - α) percentage point of the chi-square distribution with m degrees of freedom.

If m = p, then the procedure above yields a test that all the β_j^* are simultaneously zero, that is $\boldsymbol{\beta}^* = \mathbf{0}$. In this case the test statistic Q^* given in (12) becomes the weighted sum of squares due to regression

$$Q_R^* = \left(\hat{\boldsymbol{\beta}}^*\right)' \left(\boldsymbol{\Sigma}^*\right)^{-1} \hat{\boldsymbol{\beta}}^* .$$

The test that $\boldsymbol{\beta}^* = \mathbf{0}$ is simply a test of whether the weighted sum of squares due to the regression is larger than would be expected if $\boldsymbol{\beta}^* = \mathbf{0}$, and the test consists of rejecting the hypothesis that $\boldsymbol{\beta}^* = \mathbf{0}$ if Q_R^* exceeds the 100(1 - α) percentage point of a chi-square with p degrees of freedom.

Testing Whether the Between-studies Variance Component $\tau^2 = 0$

It seems reasonable that the greater the variation in the observed effect size estimates, the stronger the evidence that $\tau^2 > 0$. A simple test (the likelihood ratio test) of the hypothesis that $\tau^2 = 0$ uses the weighted sum of squares about the weighted mean that would be obtained if $\tau^2 = 0$. Specifically, it uses the statistic

$$Q = \sum_{i=1}^{k} (T_i - \hat{\beta}_0)^2 / v_i , \qquad (13)$$

where $\widehat{\beta}_0$ is the estimate of β_0 that would be obtained from equation (1) under the hypothesis that $\tau^2 = 0$. The statistic Q has the chi-squared distribution with (k – 1) degrees of freedom if $\tau^2 = 0$. Therefore a test of the null hypothesis that $\tau^2 = 0$ at significance level α rejects the hypothesis if Q exceeds the 100(1 – α) percent point of the chi-square distribution with (k – 1) degrees of freedom.

This (or any other statistical hypothesis test) should not be interpreted too literally. The test is not very powerful if the number of studies is small or if the conditional variances (the v_i) are large (see Hedges and Pigott, 2001). Consequently, even if the test does not reject the hypothesis that $\tau^2 = 0$, the actual variation in effects across studies may be consistent with a substantial range on nonzero values of τ^2, some of them rather large. That is it is unlikely that the between-study variance is *exactly* zero. This suggests that it is important to consider estimation of τ^2 and use these estimates in constructing estimates of the mean.

Estimating the Between-studies Variance Component τ^2

Estimation of τ^2 can be accomplished without making assumptions about the distribution of the random effects or under various assumptions about the distribution of the random effects using other methods such as maximum likelihood estimation. Maximum likelihood estimation is more efficient if the distributional assumptions about the study-specific random effects are correct, but these assumptions are often difficult to justify theoretically and difficult to verify empirically. Thus distribution free estimates of the between-studies variance component are often attractive.

A simple, distribution free estimate of τ^2 is given by

$$\hat{\tau}^2 = \begin{cases} \dfrac{Q-(k-1)}{a} & if\ Q \geq (k-1) \\ 0 & if\ Q < (k-1) \end{cases}$$

where a is given by

$$a = \sum_{j=1}^{k} w_i - \frac{\sum_{j=1}^{k} w_i^2}{\sum_{j=1}^{k} w_i},$$

and $w_i = 1/v$ and Q is defined in (13). Estimates of τ^2 are set to 0 when Q - (k - 1) yields a negative value, since τ^2, by definition, cannot be negative.

Testing the Significance of the Residual Variance Component

It is sometimes useful to test the statistical significance of the residual variance component τ^2 in addition to estimating it. The test statistic used is

$$Q_E = \mathbf{T}'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{T},$$

where $\mathbf{V} = \text{Diag}(v_1, \ldots v_k)$. If the null hypothesis

$$H_0: \tau^2 = 0$$

is true, then the weighted residual sum of squares Q_E has a chi-square distribution with k - p degrees of freedom (where p is the total number of predictors including the intercept). Therefore the test of H_0 at level α is to reject if Q_E exceeds the $100(1 - \alpha)$ percent point of the chi-square distribution with (k - p) degrees of freedom.

Example

We return to our example of the studies of gender differences in field articulation ability (data presented in Table 1). First we turn to the question of whether the effect sizes have more sampling variation than would be expected from the size of their conditional variances. Computing the test statistic Q we obtain Q = 24.103, which is slightly larger than 22.36, which is the 100(1 –

0.05) = 95 percent point of the chi-square distribution with 14 – 1 = 13 degrees of freedom. Actually a Q value of 24.103 would occur only about 3% of the time if $\tau^2 = 0$. Thus there is some evidence that the variation in effects across studies is not simply due to chance sampling variation.

The next step is to investigate how much variation there might be across studies. Hence, we compute the estimate of τ^2 (the variation of effect sizes estimates across studies) using the distribution free method described above. We obtain the estimate

$$\hat{\tau}^2 = \frac{24.103 - (14-1)}{195.384} = 0.057 .$$

Notice that this value of $\hat{\tau}^2$ is about 65% of the average sampling error variance. This indicates that the between-study variation is not negligible in this sample.

Now, we compute the weighted mean of the effect size estimates. In this case the weights include the estimate of $\hat{\tau}^2$. This yields an overall mean estimate of $\hat{\beta}_0^* = 0.549$ with a variance of $v_\bullet^* = 0.0094$. Notice that the variance of the weighted mean is now two times as large as in the fixed effects case. The 95% confidence interval for β_0^* is given by $0.3396 = 0.549 - 2.160\sqrt{0.0094} \leq \beta_0^* \leq 0.549 - 2.160\sqrt{0.0094} = 0.7584$. This confidence interval does not include 0, so the data are incompatible with the hypothesis that $\beta_0^* = 0$. Alternatively, the ratio $\hat{\beta}_0^* / v_\bullet^* = 5.67$ which indicates that the overall mean is significantly different from zero since the observed value is larger than the two-tailed critical t-value with 13 degrees of freedom at the $\alpha = 0.05$ significance level (2.160).

Now consider the case where the year of study is entered in the regression equation. Since the year of study will explain between-study variation we need to compute the residual estimate of

$\hat{\tau}^2$.

The distribution-free method of the estimation involves computing an estimate of the residual variance component and then computing a weighted least squares analysis conditional on this variance component estimate. Whereas the estimates are "distribution-free" in the sense that they do not depend on the form of the distribution of the random effects, the tests and confidence statements associated with these methods are only strictly true if the random effects are normally distributed. The usual estimator is based on the statistic used to test the significance of the residual variance component. It is the natural generalization of the estimate of between-study variance component given for example by Dersimonian and Laird (1986). Specifically, the usual estimator of the residual variance component is given by

$$\hat{\tau}^2 = (Q_E - k + p)/c$$

where Q_E is the test statistic used to test whether the residual variance component is zero (the residual sum of squares from the weighted regression using weights $w_i = 1/v_i$ for each study)

and c is a constant given by

$$c = tr(\mathbf{V^{-1}}) - tr[(\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-2}X}],$$

where $\mathbf{V} = diag(v_1, ..., v_k)$ is a k x k diagonal matrix of conditional variances and tr(**A**) is the trace of the matrix **A**.

First we compute the constant c as c = 174.537 and the Q_E as $Q_E = 15.11$. Hence, $\hat{\tau}^2 = (15.11 - 12)/174.537 = 0.0178$ which is three times smaller now). This value of $\hat{\tau}^2$ is now incorporated in the weights and the computation of the regression coefficients. The estimated regression coefficients are $\hat{\beta}_0 * = 3.217$ for the intercept term and $\hat{\beta}_1 * = -0.040$ for the effect of year. The variances of the

regression estimates are 1.25633 for the intercept term and 0.00028 for the year of study effect. The 95% confidence interval for β_1^* is given by $-0.0765 = -0.040 - 2.179\sqrt{0.00028} \leq \beta_1^* \leq -0.040 + 2.179\sqrt{0.00028} = -0.0035$. This confidence interval does not include 0, so the data are incompatible with the hypothesis that $\beta_1^* = 0$. Alternatively, the ratio $\hat{\beta}_1^* / \text{var}\left(\hat{\beta}_1^*\right) = -2.386$ which indicates that year effect is significantly different from zero since the absolute observed value is larger than the two-tailed critical t-value at the $\alpha = 0.05$ significance level with 12 degrees of freedom (2.179). This indicates that the effect size estimates get smaller over time (as in the fixed effects analyses). Again, the above results are easily obtained using the second version of comprehensive meta-analysis by Hedges et al., (2005).

Multivariate Meta-analysis

In the previous sections, we developed methods for fitting general linear models to the effect sizes from a series of studies when the effect size estimates are independent. In this section we sketch analogues to those methods when the sampling errors are not independent. These methods are essentially multivariate generalizations of the fixed and mixed effects models given above for univariate meta-analys1s. To use these methods, the joint distribution of the non-independent effect size estimates must be known, which typically involves knowing both the variances and the covariance structure of the effect size estimates. The sampling distribution of correlated effect size estimates is discussed by Gleser and Olkin (1994).

Fixed Effects Models for Correlated Effect Size Estimates

A researcher may be interested in fixed effects models for the analysis of the relation

between study characteristics (study-level covariates) and effect sizes. In fixed effects models the effect size parameter is assumed to be fixed at a certain value. The only source of variation in such models is the sampling variation due to different samples of individuals. As in the univariate case, natural tests of goodness of fit are provided for the fixed effects analysis. They test the hypothesis that the variability among studies is no greater than would be expected if all of the variation among effect size parameters is explained by the linear model. These tests are generalizations of the test of homogeneity of effect size and the tests of goodness of fit for linear models given previously.

Assume that each $\mathbf{T}_i$ has a q-variate normal distribution (since there may be q effect size estimates in each study) about the corresponding $\boldsymbol{\theta}_i$ with known q x q covariance matrix $\boldsymbol{\Sigma}_i$, that is

$$\mathbf{T}_i \sim N(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i),\ i = 1, ..., k. \qquad (14)$$

There is no need for all studies to have the same number of effect sizes, but we make that assumption here to simplify notation. We denote the kq dimensional column vectors of population and sample effect sizes by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', ..., \boldsymbol{\theta}_k')'$ and $\mathbf{T} = (\mathbf{T}_1', ..., \mathbf{T}_k')'$, respectively, where $\mathbf{T}_i = (T_{i1}, ..., T_{iq})'$ and $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{iq})'$

In the present model assume that the effect size parameter vector $\boldsymbol{\theta}_i$ for the i^{th} study depends on a vector of p fixed study-level covariates $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$. Specifically, assume that

$$\theta_{ij} = \beta_{1j}x_{i1} + \cdots + \beta_{pj}x_{ip},\ i = 1, ..., k;\ j = 1, ..., q \qquad (15)$$

where $\beta_{1j}, ..., \beta_{pj}$ are unknown regression coefficients in the model for the jth component of $\boldsymbol{\theta}_i$. In order to utilize the standard form of generalized least squares, it is helpful to stack the elements of $\mathbf{T}_i$, $\boldsymbol{\theta}_i$, and $\boldsymbol{\beta}_j$ so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', ..., \boldsymbol{\theta}_k')' = \mathbf{1}_k \otimes \boldsymbol{\theta}_i$, $\mathbf{T} = (\mathbf{T}_1', ..., \mathbf{T}_k')' = \mathbf{1}_k \otimes \mathbf{T}_i$, and $\boldsymbol{\beta} = (\beta_{11}, ..., \beta_{p1}, \beta_{21}, ..., \beta_{pq})'$ are column vectors, $\mathbf{1}$ is a column vector of ones and $\otimes$ is the Kronecker product operator (and i = 1,

…, k).

In the univariate case the design matrix would be

$$\mathbf{X}_U = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)'$$

where $\mathbf{x}_i = (x_{i1},...,x_{ip})'$. In the multivariate case however, each $\mathbf{x}_i$ must be repeated q times for study i (once for each outcome in the study), so that the multivariate design matrix is

$$\mathbf{X} = (\mathbf{I}_q \otimes \mathbf{x}_1, \mathbf{I}_q \otimes \mathbf{x}_2, ..., \mathbf{I}_q \otimes \mathbf{x}_k)' \quad (16)$$

where $\mathbf{I}_q$ is a q x q identity matrix and $\otimes$ is the Kronecker product operator. The design matrix $\mathbf{X}$ has dimension kq x pq and is assumed to have no linearly dependent columns; that is, $\mathbf{X}$ has rank p. Equation (15) can be written succinctly in matrix notation as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_{11},...,\beta_{p1}, \beta_{21},...,\beta_{pq})'$ is the pq-dimensional vector of regression coefficients.

Then the model for $\mathbf{T}$ can be written as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (17)$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_{kq})' = \mathbf{T} - \boldsymbol{\theta}$ is a kq-dimensional column vector of residuals.

Weighted Least Squares Estimator of the Regression Coefficients

The linear model $\mathbf{T} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for the effect sizes is analogous to the model that is the basis for ordinary least squares regression analysis. Because $\boldsymbol{\varepsilon} = \mathbf{T} - \boldsymbol{\theta}$, it follows that the distribution of $\boldsymbol{\varepsilon}$ is kq-variate normal with mean zero and known kq x kq block-diagonal covariance matrix $\mathbf{V}$ given by

$$\mathbf{V} = \text{Diag}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\Sigma}_k) = \mathbf{I}_k \otimes \boldsymbol{\Sigma}_i \qquad i = 1, \ldots, k \quad (18)$$

Thus, the elements of $\boldsymbol{\varepsilon}$ are not identically or independently distributed, but the covariance matrix is known. Therefore we can use the method of generalized least squares to obtain an estimate of $\boldsymbol{\beta}$.

This is essentially the approach of Raudenbush, Becker, and Kalaian (1988), Gleser and Olkin (1994), and Berkey, Anderson, and Hoaglin (1996).

The generalized least squares estimator $\hat{\boldsymbol{\beta}}$ under the model (17) with covariance matrix V given in (18), which is also the maximum likelihood estimator of **β** under that model, is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{T}, \tag{19}$$

which has a pq-variate normal distribution with mean **β** and covariance matrix **Σ** given by

$$\boldsymbol{\Sigma} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{20}$$

Tests and Confidence Intervals

Once the estimate of **β** and its covariance matrix **Σ** are calculated, tests and confidence intervals for individual regression coefficients and tests for blocks of regression coefficients correspond exactly to those in the univariate fixed effects model described in previously. Tests of goodness of fit of regression models are straightforward generalizations of those used in the univariate general linear model using the estimate $\hat{\boldsymbol{\beta}}$ and **Σ**.

Example: Studies of the Effects of Coaching on the SAT

A collection of 19 studies of the effects of coaching on SAT verbal and SAT math scores was assembled by Kalaian and Raudenbush (1996). They examined the question of whether the effects of coaching were greater if the length of coaching was greater. The study level covariate was the log of the number of hours spent in coaching classes. The data are presented in Table 2. We compute the estimate of the regression coefficients from equation (19) as $\hat{\boldsymbol{\beta}}$ = (-0.12948, 0.07900, -0.2936, 0.1349)′. We also compute the standard errors of the coefficients of $\hat{\boldsymbol{\beta}}$ as the square roots of the

diagonal elements of equation (20) as SE($\hat{\beta}_1$) = $\sqrt{\sigma_{11}}$ = 0.2186, SE($\hat{\beta}_2$) = $\sqrt{\sigma_{22}}$ = 0.0703, SE($\hat{\beta}_3$) = $\sqrt{\sigma_{33}}$ = 0.2194, and SE($\hat{\beta}_4$) = $\sqrt{\sigma_{44}}$ = 0.0705. Computing the individual test statistics for the four regression coefficients we obtain t_1 = -0.592, t_2 = 1.124, t_3 = -1.338, t_4 = 1.913. Notice that none of two-tailed tests is statistically significant at the 0.05 significance level. However, t_4 is significant at the 0.5 level and if we assume a one-tailed test.

Mixed Models for Correlated Effect Size Estimates

When there is nonnegligible covariation among effect size parameters even after controlling for the factors that are of interest in the analysis, a general linear model analysis of effect size data is more appropriate. That is, there is greater residual covariation than would be expected from sampling error alone. The mixed model incorporates a component of between-study covariation into the uncertainty of effect size parameters and their estimates which has the effect of increasing residual variation.

Assume that each $\mathbf{T}_i$ has a q_i-variate normal distribution about the corresponding $\boldsymbol{\theta}_i$ with known q_i x q_i covariance matrix $\boldsymbol{\Sigma}_i$, that is

$\mathbf{T}_i \sim N(\boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i)$, i = 1, ..., k.

To simplify notation, we require all of the studies have the same number of effect sizes. We denote the kq dimensional column vectors of population and sample effect sizes by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', ..., \boldsymbol{\theta}_k')'$ and $\mathbf{T} = (\mathbf{T}_1', ..., \mathbf{T}_k')'$, respectively. This model is identical to the fixed effects model.

In the present model assume that the effect size parameter vector $\boldsymbol{\theta}_i$ for the ith study depends on a vector of p fixed predictor variables $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$. Specifically, assume that

$$\theta_{ij} = \beta_{1j} x_{i1} + \cdots + \beta_{pj} x_{ip} + \xi_{ij} \quad i = 1, ..., k, \; j = 1, ..., q \qquad (21)$$

where $\beta_{11},...,\beta_{pq}$ are unknown regression coefficients, and ξ_{ij} is a random effect.

As in the fixed effects case, the design matrix **X** is assumed to have no linearly dependent columns; that is, **X** has rank p. As in the fixed effects case the dimension of the design matrix **X** is kq x pq. Equation (210 can be written succinctly in matrix notation as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{I}\boldsymbol{\Xi} = (\mathbf{I}_q \otimes \mathbf{x}_1, \mathbf{I}_q \otimes \mathbf{x}_2, ..., \mathbf{I}_q \otimes \mathbf{x}_k)'\boldsymbol{\beta} + \mathbf{I}\boldsymbol{\Xi} \quad (22)$$

where $\boldsymbol{\beta} = (\beta_1,...,\beta_{pq})'$ is the pq dimensional vector of regression coefficients, **I** is a kq dimensional identity matrix, $\boldsymbol{\Xi}$ is a kq dimensional vector of between-study random effects, $\mathbf{I}_q$ is a q x q identity matrix, and $\otimes$ is the Kronecker product operator. The vector $\boldsymbol{\Xi}$ of the between-study random effects follows a q-variate normal with mean zero and q x q covariance matrix $\boldsymbol{\Omega}$ of the between-study variance components.

The model for **T** can then be written as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{I}\boldsymbol{\Xi} + \boldsymbol{\varepsilon}, \quad (23)$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_{kq})' = \mathbf{T} - \boldsymbol{\theta}$ is a kq dimensional column vector of residuals.

Estimation of the Regression Coefficients and the Covariance Components

The regression coefficient vector $\boldsymbol{\beta}$ and the covariance component matrix $\boldsymbol{\Omega}$ can be estimated by weighted least squares as in the case of the univariate mixed model. The usual procedure is to first estimate the covariance component matrix $\boldsymbol{\Omega}$ and then reweight to estimate the regression coefficient vector $\boldsymbol{\beta}$ and its covariance matrix $\boldsymbol{\Sigma}$. There are usually advantages (among them software availability) in considering the problem as a special case of the hierarchical linear model considered in the previous section in conjunction with univariate mixed model analyses. The multivariate mixed model analyses can be carried out as instances of the multivariate hierarchical linear model (see Thum, 1997), estimating parameters by the method of by maximum likelihood. However a simpler

alternative is available.

Since the sampling error covariance matrix is known, it is possible to transform the within-study model so that the sampling errors are independent (Kalaian and Raudenbush, 1996). For each study perform the Cholesky factorization of each sampling error covariance matrix $\mathbf{\Sigma}_i$ so that

$$\mathbf{\Sigma}_i = \mathbf{F}_i\mathbf{F}_i',$$

where $\mathbf{F}_i$ is a known matrix (since $\mathbf{\Sigma}_i$ is a known matrix) and is the lower triangular (square root) matrix of the Cholesky decomposition. Then transform the within-study model to be

$$\mathbf{F}_i^{-1}\mathbf{T}_i = \mathbf{F}_i^{-1}\boldsymbol{\theta}_i + \mathbf{F}_i^{-1}\boldsymbol{\varepsilon}_i.$$

The transformed effect size vector $\mathbf{Z}_i$ given by

$$\mathbf{Z}_i \equiv \mathbf{F}_i^{-1}\mathbf{T}_i$$

has a sampling error vector

$$\tilde{\boldsymbol{\varepsilon}}_i = \mathbf{F}_i^{-1}\boldsymbol{\varepsilon}_i$$

which has covariance matrix $\mathbf{I}$, a q_i x q_i identity matrix. Thus one might write the within-study model as

$$\mathbf{Z}_i = \mathbf{F}_i^{-1}\boldsymbol{\theta}_i + \tilde{\boldsymbol{\varepsilon}}_i, \qquad (24)$$

where the transformed effect size estimates $\mathbf{Z}_i$ are now independent with a constant variance, but the effect size parameter vector $\boldsymbol{\theta}_i$ is the same as in the original model.

Thus the within-study model in (24) along with the between-study model is now a conventional two-level linear model with independent sampling errors at the first level. Therefore, conventional software can be used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ by the method of maximum likelihood (as in HLM).

Multivariate Meta-analysis Using HLM

HLM is a software package designed especially for fitting multi-level models, and it can be used to fit mixed effects models to effect size data with study level covariates (Raudenbush, Bryk, Cheong, and Congdon, 2004). It can also be used to fit multivariate mixed models to effect size data in meta-analysis. Table 3 describes the input file for a mixed model multivariate meta-analysis of the SAT coaching data reported by Kalaian and Raudenbush (1996). The data for the analysis are read from a separate file and consist of 19 pairs of effect sizes from 19 studies of the effects of coaching on the SAT verbal and SAT math tests. The first three lines set the maximum number of iterations the program will run (NUMIT:1000), the criterion for stopping iteration (STOPVAL:0.0000010000), and that a linear model will be used (NONLIN: n). Lines four to six indicate the level I model (LEVEL 1: MATH=VERBAL+MATH+RANDOM), and the level II models (LEVEL 2: VERBAL = INTRCPT2 + HOURS+RANDOM/ and LEVEL 2: MATH=INTRCPT2+HOURS+RANDOM/). Lines seven and eight indicate that no weights are used in the computations (LEVELWEIGHT:NONE). Line nine indicates that the variance is not known (VARIANCEKNOWN:NONE), line 10 that no output file of residuals is requested (RESFIL:N), and line 11 that the level 1 variances are not heterogeneous (HETEROL1VAR: n). Line 12 indicates that the default value of the accelerator should be used in estimation (ACCEL:5), line 13 that a latent variable regression is not used (LVR:N), and line 14 that the OL equations should be printed to 19 units (LEV1OLS:10). Line 15 indicates that restricted maximum likelihood is used (MLF:N), line 16 that no optional hypothesis testing will be done (HYPOTH:N), and line 17 that unacceptable starting values of τ will be automatically corrected (FIXTAU:3). Line 18 indicates that none of the fixed effects are constrained to be equal to one another (CONSTRAIN:N). Line 19 specifies that the output file is named "COACHING.OUT", line

20 that the full output will be given (FULLOUTPUT:Y), and line 21 specifies the title of the output.

The results are reported in Table 4. The top panel of Table 4 shows the regression coefficients estimates. The estimates are only slightly different than those in the fixed effects analyses. Overall, as in the fixed effects analyses most of the regression estimates are not significantly different from zero (except for hours of coaching). The predictor hours of coaching is significant in verbal, indicating that hours of coaching matters in verbal. The bottom panel of Table 4 shows the variance component estimates for the residuals about the SAT verbal and SAT math regressions, respectively, along with the Chi-square test of the hypothesis that the variance component is zero and the p-value for that test. Both variance components for SAT math and verbal are not significantly different from zero indicating that there is negligible between-study variation.

Conclusion

This study presented univariate and multivariate models for meta-analysis. The use of fixed and mixed effects models in univariate and multivariate settings was also demonstrated. Specialized statistical software packages such as comprehensive meta-analysis can be easily used to conduct univariate weighted least-squares analyses in meta-analysis (both for fixed and mixed effects analyses). Other specialized software packages such as HLM can carry out multivariate mixed models analyses for meta-analytic data with nested structure. Mixed effects models analyses can also be performed with specialized software such as MLwin and the SAS procedure proc mixed. The mixed effects models presented here can be extended to three or more levels of hierarchy capturing random variation at higher levels. For example, a three-level meta-analysis can model and compute variation between investigators or laboratories at the third level (Konstantopoulos, 2005).

Endnote

[1] This software is especially designed to cover various methods for meta-analytic data (www.Meta-Analysis.com).

Bibliography

Berkey, C. S., Anderson, J. J., & Hoaglin, D. C. (1996). Multiple-outcome meta-analysis of clinical trials. *Statistics in Medicine, 15*, 537-557.

Cohn, L. D., & Becker, B. J. (2003). Title: How meta-analysis increases statistical power *Psychological Methods, 8*, 243-253.

Cooper, H. (1989). *Integrating research (Second Edition)*. Newbury Park, CA: Sage Publications.

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

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

Fleiss, J. L. (1994). Measures of effect size for categorical data. Pages 245-260 in H. Cooper and L. V. Hedges, *The handbook of research synthesis.* New York: The Russell Sage Foundation.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. Pages 339-356 in H. Cooper & L. V. Hedges (Eds.) *The handbook of research synthesis*. New York: The Russell Sage Foundation.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*, 490-499.

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

Hedges, L. V. (1994). Fixed effects models. Pages 285-299 in H. Cooper and L. V. Hedges, *The handbook of research synthesis.* New York: The Russell Sage Foundation.

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical test in meta-analysis. *Psychological Methods, 6,* 203-217.

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

Hedges, L. V., Borenstein, M., Higgings, J., & Rothstein, H. (2005). *Comprehensive meta-analysis*. Englewood, NJ: Biostat.

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

Kalaian, H. & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1*, 227-235.

Konstantopoulos, S. (2005). *Three-level models in meta-analysis*. Paper presented at the annual American Educational Association Conference, Montreal.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences. Stanford, CA: Stanford University press.

Raudenbush, S. W., Becker, B. J., & Kalaian, S. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103*, 111-120.

Raudenbush, S. W., Bryk, A., Cheong, Y. F., & Congdon, R. (2005). *HLM 6: Hierarchical linear and onlinear modeling*. Lincolnwood, IL: Scientific Software International.

Rosenthal, R. (1994). Parametric measures of effect size. Pages 231-244 in H. Cooper and L. V. Hedges, *The handbook of research synthesis*. New York: The Russell Sage Foundation.

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

Schmidt, F. L. & Hunter, J. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi -experimentaldesigns for generalized causal inference*. Boston, MA: Houghton Mifflin.

Smith, M. l. & Glass, G. V (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760.

Thum, Y. M. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics, 22*, 77-108.

Table 1

Field Articulation Data from Hyde (1981).

ID	ES	Var	Year
1	0.76	0.071	1955
2	1.15	0.033	1959
3	0.48	0.137	1967
4	0.29	0.135	1967
5	0.65	0.140	1967
6	0.84	0.095	1967
7	0.70	0.106	1967
8	0.50	0.121	1967
9	0.18	0.053	1967
10	0.17	0.025	1968
11	0.77	0.044	1970
12	0.27	0.092	1970
13	0.40	0.052	1971
14	0.45	0.095	1972

Note: ID = Study ID; ES = Effect Size Estimate; VAR = Variance; Year = Year of Study.

Table 2

SAT Coaching Data from Kalaian and Raudenbush (1996): Selected Sample.

ID	SAT(V)	SAT(M)	VAR(V)	COV(V,M)	VAR(M)	Log(Hours)	Year
9	0.13	0.12	0.01468	0.00968	0.01467	3.044522438	73
10	0.25	0.06	0.02180	0.01430	0.02165	3.044522438	73
11	0.31	0.09	0.02208	0.01444	0.02186	3.044522438	73
12	0.00	0.07	0.14835	0.09791	0.14844	2.186051277	86
26	0.13	0.48	0.12158	0.08049	0.12481	3.178053830	88
29	-0.23	0.33	0.25165	0.16397	0.25340	2.890371758	87
30	0.13	0.13	0.09327	0.06151	0.09327	2.708050201	85
31	0.13	0.34	0.04454	0.02944	0.04509	3.401197382	60
33	0.09	-0.11	0.03850	0.02536	0.03852	2.302585093	62
34	-0.10	-0.08	0.10657	0.07030	0.10653	1.791759469	88
35	-0.14	-0.29	0.10073	0.06654	0.10152	1.791759469	88
36	-0.16	-0.34	0.10917	0.07214	0.11039	1.791759469	88
37	-0.07	-0.06	0.10889	0.07185	0.10887	1.791759469	88
38	-0.02	0.21	0.01857	0.01225	0.01861	2.708050201	58
39	0.06	0.17	0.00963	0.00636	0.00966	2.708050201	53
42	0.15	0.03	0.00668	0.00440	0.00667	3.688879454	78
43	0.17	-0.19	0.10285	0.06748	0.10294	2.639057330	76
45	-0.04	0.60	0.03203	0.02110	0.03331	4.143134726	87
47	0.54	0.57	0.07968	0.05206	0.07998	3.295836866	88

Note: ID = Study ID; SAT = Scholastic Aptitude Test; V = Verbal; M = Math; VAR = Variance; COV = Covariance.

Table 3

HLM Input for Mixed Model Multivariate Analyses of SAT Coaching Data from Kalaian and Raudenbush (1996)

Input File

```
NUMIT:1000
STOPVAL:0.0000010000
NONLIN:n
LEVEL1:MATH=VERBAL+MATH+RANDOM
LEVEL2:VERBAL=INTRCPT2+HOURS+RANDOM/
LEVEL2:MATH=INTRCPT2+HOURS+RANDOM/
LEVEL1WEIGHT:NONE
LEVEL2WEIGHT:NONE
VARIANCEKNOWN:NONE
RESFIL2:N
HETEROL1VAR:n
ACCEL:5
LVR:N
LEV1OLS:10
MLF:n
HYPOTH:n
FIXTAU:3
CONSTRAIN:N
OUTPUT:COACHING.OUT
FULLOUTPUT:Y
TITLE:MULTIVARIATE META ANALYSIS USING HLM
```

Table 4

HLM Output for Mixed Model Multivariate Analyses of SAT Coaching Data from Kalaian and Raudenbush (1996)

Output File

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For VERBAL, B1					
INTRCPT2, G10	-0.051329	0.227003	-0.226	17	0.824
HOURS, G11	0.049071	0.073447	0.668	17	0.513
For MATH, B2					
INTRCPT2, G20	-0.496924	0.264238	-1.881	17	0.077
HOURS, G21	0.212755	0.087375	2.435	17	0.026

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
VERBAL,	U1	0.05144	0.00265	17	8.80514	>.500
MATH,	U2	0.12414	0.01541	17	18.40913	0.363