Best practices in quantitative methods

12 An Introduction to Meta–Analysis

Contributors: Jason Osborne
Print Pub. Date: 2008
Online Pub. Date:
Print ISBN: 9781412940658
Online ISBN: 9781412995627
DOI: 10.4135/9781412995627
Print pages: 177-195

10.4135/9781412995627.d15

[p. 177 ↓]

12 An Introduction to Meta–Analysis

Spyros Konstantopoulos

Researchers in the social sciences often already have access to completed studies in the literature that relate to or address their hypotheses. How best, then, to organize and summarize findings from these studies in order to identify and exploit what is known and focus research on promising areas? While narrative summaries and analyses of the literature are important (and the norm), quantitative research synthesis or meta–analysis is currently considered a best practice across many disciplines (see Cooper & Hedges, 1994; Hedges & Olkin, 1985).

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, such as the effect of some treatment or intervention, where the average effect of the treatment is often of interest (see, e.g., Smith & Glass, 1977). In other cases, the degree of variation in results across studies will be of primary interest, where meta–analysis can be used to study the generalizability of employment test validities across situations (see, e.g., Schmidt & Hunter, 1977). Meta–analysis is

also frequently used to identify the contexts in which a treatment or intervention is most successful or has the largest effect (see, e.g., Cooper, 1989).

Meta–analytic reviews are designed to integrate empirical research with the objective to create research generalizations; hence, one substantial advantage of meta–analysis is the generality of the summary estimates (Cooper & Hedges, 1994). This constitutes a unique aspect of meta–analysis that is crucial for the external validation of the estimates (see Shadish, Cook, & Campbell, 2002). Generally, the estimates that are produced from meta–analyses have higher **[p. 178 ↓]** external validity than estimates reported in single studies. Other advantages of meta–analytic reviews include the fact that the summary estimates that are generated from such reviews can support or refute theories (and hence facilitate the improvement of substantive theory) and can guide future research by identifying important issues (Cooper, 1989). In addition, from a statistical point of view, the results of meta–analytic procedures have higher statistical power than do indicators obtained from individual studies, which increases the probability of detecting associations of interest (Cohn & Becker, 2003).

The term *meta–analysis* is sometimes used to describe the entire process of research synthesis or integrative research review. However, more recently, it has been used specifically for the statistical component of research synthesis (Cooper, 1989; Cooper & Hedges, 1994). Other components of research synthesis that take place prior to meta–analysis include the formulation of the question of interest (or problem), the search of the literature or data collection, and the evaluation and coding of the data that involve the evaluation of the quality of the data and the creation of variables and quantitative indexes (see Cooper, 1989). 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 poor–quality data. Excellent treatments of the nonstatistical aspects of research synthesis are available in Cooper (1989), Cooper and Hedges (1994), and Lipsey and Wilson (2001).

Early Stages of Research Synthesis

In the very early stages of meta–analytic reviews, the reviewers need to clearly formulate a question of interest and familiarize themselves with what theorists and

empirical researchers have discussed on that specific topic. The next step involves constructing a coding sheet to record important information from the sample of studies collected. The coding sheet can include general information about the authors of the study, the year of publication of the study, the source of the study (e.g., journal title), the research design used in the study (e.g., correlational or experimental), the characteristics of the individuals who participated in the study (e.g., age, gender, numbers of participants), and the outcome measures of the study. Most important for meta–analysis, however, the coding sheet should include information about the summary statistics of the study. In the social sciences, these statistical outcomes typically include means, standard deviations, and sample sizes (for groups of individuals); correlation coefficients; odds ratios; and the value of the test (e.g., *t* test) and the sample or the *p* value of the test and the sample size. Cooper (1989) provides a thorough discussion about coding sheets. Recently, software packages such as Comprehensive Meta–Analysis (CMA), which are designed especially to conduct meta–analysis, have offered multiple formats for entering meta–analytic data.[1]

The next stage involves the literature review in order to locate the relevant studies. In this stage, it is important that the meta–analyst use multiple sources of literature retrieval in order to ensure that useful studies that are related directly to the question of interest are included in the sample (see White, 1994). Common ways of conducting literature searches include tracing references in previous relevant review studies, references in relevant books, references in nonreview relevant studies from journals that researchers subscribe to, references through computer searches of relevant databases (e.g., web of science, ERIC, PsycINFO, Econ Lit, Sociological Abstracts, Dissertation Abstracts, etc.), and references through a manual search of journals that typically publish work on the specific topic. In addition, informal channels of locating studies include communication with researchers who work or have worked on the specific topic and informal conversations with other researchers or students in conferences (see Cooper, 1989; White, 1994).

It is important at this stage that the sample of studies includes published and unpublished work so that the sample represents accurately the number of studies that were actually undertaken. This indicates that the inclusion of a study in the sample should not depend on the statistical significance of the results but on the relevance

of the study. Specifically, if the sample of studies includes only published work, it is possible that the largest or more significant effect size estimates are overrepresented in the sample since significant results (or larger estimates) are more likely to be published. Hence, the estimates derived from published work form a selected **[p. 179 ↓]** subsample, and this can lead to selection or publication bias. There are several ways to examine publication bias. A common way to examine publication bias is the funnel plot that plots the sample size versus the effect size for each study (see Light & Pillemer, 1984). When the graph resembles a funnel, publication bias seems unlikely. Another way to examine publication bias is through a z test (see Begg, 1994). In this case, when the z test is statistically significant, there is evidence of publication bias. Rosenthal's (1979) fail–safe (or file drawer) method is another well–known technique that computes the number of missing studies (with a mean effect of zero) that would need to be added to the analysis to yield a statistically insignificant overall effect. Large numbers of missing studies would indicate that publication bias is rather unlikely. A recent method, called trim and fill, also accounts for publication bias by imputing the missing studies, adding them to the analysis, and recomputing an overall effect size (Duval & Tweedie, 2000). A thorough discussion about publication bias in meta–analysis is provided by Rothstein, Sutton, and Borenstein (2005). Software packages such as CMA provide multiple methods that examine publication bias and assess its impact on the summary estimates.

Finally, at the evaluation stage, the reviewer needs to make critical judgments about the quality of the data and create consistent and objective criteria for including studies in the sample (Cooper, 1989; Wortman, 1994). According to Cooper (1989), the validity of the study's methods is a crucial criterion for discarding or including data. That is, the reviewer needs to evaluate whether the study was conducted in a way that secures the validity of its estimates. Sometimes, reviewers decide to include a study in the sample (or exclude it). Other times, the quality of the study can be represented in a continuous scale and can be used to weight studies according to their quality (i.e., higher quality studies are assigned higher weights). Notice that a weight of zero is equivalent to excluding a study. Shadish and Haddock (1994) demonstrate how weights that indicate the quality of the study can be incorporated in the computation of meta–analytic summary estimates. Of course, the inclusion of the study also depends on whether the study provides the required information for computing estimates related to

the question of interest of the review. A thorough discussion about data evaluation is provided by Cooper (1989) and Wortman (1994).

Meta–Analysis

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 estimates will vary somewhat from study to study (sampling variation), while the effect size parameter is in principle fixed (fixed effects models). 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, the statistical analysis used in each study, and the information provided 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

$$d = \frac{\overline{Y}^T - \overline{Y}^C}{S}, \qquad (1)$$

[p. 180 ↓] where Y^T is the sample mean of the outcome in the treatment group, 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

$$\delta = \frac{\mu^T - \mu^C}{\sigma}, \qquad (2)$$

where μ^T is the population mean in the treatment group,is the population mean outcome in the control group, and τ 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 & Olkin, 1985). The sampling uncertainty of the standardized mean difference is characterized by its variance, which is

$$v = \frac{n^T + n^C}{n^T n^C} + \frac{d^2}{2(n^T + n^C)}, \qquad (3)$$

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

$$d - 2\sqrt{v} \leq \delta \leq d + 2\sqrt{v}. \qquad (4)$$

Several variations of the standardized mean difference are also sometimes used as effect sizes (see Rosenthal, 1994). A standardized mean difference can easily be computed so long as a study reports sufficient information for its computation (e.g., means, standard deviation, sample sizes, the value and *p* value of the test, etc.).

The Log–Odds Ratio

In many studies of the effects of a treatment or intervention that measures 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

$$\begin{aligned} \log(OR) &= \log\left(\frac{p^T/(1-p^T)}{p^C/(1-p^C)}\right) \\ &= \log\left(\frac{p^T(1-p^C)}{p^C(1-p^T)}\right), \end{aligned} \qquad (5)$$

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

$$\omega = \log\left(\frac{\pi^T/(1-\pi^T)}{\pi^C/(1-\pi^C)}\right) = \log\left(\frac{\pi^T(1-\pi^C)}{\pi^C(1-\pi^T)}\right), \quad (6)$$

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

$$v = \frac{1}{n^T p^T} + \frac{1}{n^T(1-p^T)} + \frac{1}{n^C p^C} + \frac{1}{n^C(1-p^C)}, \quad (7)$$

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 **[p. 181 ↓]** ω. Specifically, a 95% confidence interval for the effect size is given by

$$d - 2\sqrt{v} \le \omega \le d + 2\sqrt{v}. \quad (8)$$

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 odds ratio family of effect sizes, see Fleiss (1994). Odds ratios are often reported in studies in medicine and the health sciences.

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,

$$z = \tfrac{1}{2}\log\left(\frac{1+r}{1-r}\right), \qquad (9)$$

in carrying out statistical analyses. The corresponding correlation parameter is *p*, the population correlation, and the parameter that corresponds to the estimate z is #, the z transform of p. The advantage of this transformation is that the variance of the Fisher *z* transform is independent of the correlation coefficient and is simply a function of the sample size of the study. Specifically, the sampling uncertainty of the z–transformed correlation is characterized by its variance,

$$v = \frac{1}{n-3}, \qquad (10)$$

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. Bivariate correlations are often reported in studies in the social sciences.

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 with known variance v_i. That is, assuming k studies and one estimate per study,

$$T_i \sim N(\theta_i, v_i), i = 1, \ldots, k. \qquad (11)$$

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 untrans–formed 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). A nice feature of software packages such as CMA is that the allow for transformations from one effect size estimate to another. For example, a reviewer can enter data in CMA that initially allow the computation of a standardized mean difference. However, once the standardized effect size estimate is computed, CMA can transform this estimate to a correlation coefficient, or an odds ratio, and so on (and hence the summary estimates can be expressed in various forms).

Univariate 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 & 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 ith study, let T_i. be **[p. 182 ↓]** 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,…, T_k and their variances v_1,…, v_k. The effect size estimate T_i is modeled as the effect size parameter plus a sampling error $\#_i$. That is,

$$T_i = \theta_i + \varepsilon_i,\ \varepsilon_i \sim N(0,\ v_i). \qquad (12)$$

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 $(\theta_1;\theta_2;\ldots,\#x2026;,\ \theta_k)$ were sampled. Note that this is not conceptually the same as the mean of $\theta_1, \theta_2,\ldots, \theta_k$ 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_iS are fixed and thus

$$T_i = \beta_0 + \varepsilon_i. \qquad (13)$$

Note that in meta–analysis, the variances (the **v.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

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}, \qquad (14)$$

where *w*- = 1/ **v**. Note that this estimator corresponds to a weighted mean of the *T,* 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. of $(\beta$ *0*

is simply the reciprocal of the sum of the weights

$$v_\bullet = \left(\sum_{i=1}^{k} w_i \right)^{-1}, \qquad (15)$$

and the standard error

$SE(\hat{\beta}_0)$

of

β_0

is just the square root of *v* Under this model,

β_0

is normally distributed, so a 100(1 - ?) percent confidence interval for

β_0

is given by

$$\hat{\beta}_0 - t_{\alpha/2}\sqrt{v_\bullet} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2}\sqrt{v_\bullet}, \qquad (16)$$

where *t*
a

is the $_{100\alpha}$ percent point of the *t* distribution with *(k* – 1) degrees of freedom. Similarly, a two–sided test of the hypothesis that (
0

= 0 at significance level α uses the test statistic

$$Z = \hat{\beta}_0^* / \sqrt{v_\bullet}$$

and rejects if |Z | exceeds t
#/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).

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 2th study is modeled as

$$\theta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \; i = 1, \ldots, k, \quad (17)$$

where ßj,…, ß
p

are unknown regression coefficients that need to be estimated, and x,,…, x
ip

represent values of the *p* predictors for study *i.* Thus, the model for *T*
i

is written as

$$T_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i. \quad (18)$$

To compute the regression coefficients, we use the method of generalized least squares (see appendix).

Tests for Blocks of Regression Coefficients

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

$$H_0: \beta_1 = \ldots = \beta_m = 0. \qquad (19)$$

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 to the outcome 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 **[p. 183 ↓]** 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, nonequivalent comparison group samples, and other quasi–experimental designs, but it is treated as one conceptual variable, and its importance is tested simultaneously. To test this hypothesis, we compute the statistic

$$Q = (\hat{\beta}_1, \ldots, \hat{\beta}_m)\,(\boldsymbol{\Sigma}_{11})^{-1}\,(\hat{\beta}_1, \ldots, \hat{\beta}_m)', \qquad (20)$$

where Σ_{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 of 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. In this case, the test statistic *Q* given in (20) becomes the weighted sum of squares due to regression (see appendix).

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 scored higher than females in field articulation. The effect size estimates are reported in column 2 of Table 12.1. The variances of the effect size estimates are reported in column 3. The year the study was conducted is in column 4.

First, we compute the weighted mean of the effect size estimates. This yields an overall mean estimate of

$\hat{\beta}_0$

with a variance of v. = 0.005. The 95% confidence interval for β_0

is given by $0.40 \leq \beta_0$

< 0.69. This confidence interval does not include zero, so the data are incompatible with the hypothesis that β_0

= 0. Alternatively, the ratio

$$\hat{\beta}_0/\sqrt{v_\bullet} = 7.78,$$

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 .05 significance level with 13 degrees of freedom (2.16).

Table 12.1 Field Articulation Data From Hyde (1981)

ID	*Effect Size*	*Variance*	*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

Second, we compute the effect of the year of the study. This yields an estimate of

$\hat{\beta}_1$

= -0.04, with a variance var(

$\hat{\beta}_1$

) = 0.0002. The 95% confidence interval for β_1 is given by $-0.07 \leq \beta_1 \leq -0.01$. This confidence interval does not include 0, so the data are incompatible with the hypothesis that $\beta_1 = 0$. Alternatively, the ratio

$$\hat{\beta}_1 / \sqrt{\text{var}(\hat{\beta}_1)} = -2.83,$$

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 .05 significance level with 12 degrees of freedom (2.18). This indicates that the effect size estimates get slightly smaller over time. The above results are easily obtained from the second version of CMA, developed by Hedges, Borenstein, Higgings, and Rothstein (2005).

Univariate 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 hyperparameters (usually the mean and variance) describing this population of effect **[p. 184 ↓]** parameters (see, e.g., DerSimonian & Laird, 1986; Hedges, 1983; Schmidt & Hunter, 1977). 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 nonnegligible 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 the 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. In the mixed effects model, the effect size parameter is modeled by a mean effect size β_0^*

plus a study–specific random effect η_i

that is,

$$\theta_i = \beta_0^* + \eta_i \; \eta_i \sim N(0, \tau^2). \qquad (21)$$

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). It follows that the effect size estimate *T* is modeled as

$$T_i = \beta_0^* + \eta_i + \varepsilon_i = \beta_0^* + \xi_i, \qquad (22)$$

where ξ_i is a composite error defined by $\xi_i = \eta_i + \varepsilon_i$.

Equation 22 indicates 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 the effect size estimate T_i assuming a fixed parameter θ_i and the variance of T_i incorporating the variance of the parameter θ_i

as well. The latter is the *unconditional sampling variance* of T_i

(denoted v#). Since the sampling error $\#_i$

and the random effect η_i

are assumed to be independent, and the sample variance of η_i

is
$\hat{\tau}^2$

, it follows that the unconditional sampling variance of
T_i

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

under the model is

$$\hat{\beta}_0^* = \frac{\sum_{i=1}^{k} w_i^* T_i}{\sum_{i=1}^{k} w_i^*}, \qquad (23)$$

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. Also, note that the estimate of the between–study variance is close to zero or very small, so the estimates of the mixed effects model will be similar to those obtain from a fixed effects model.

The sampling variance v.* of
$\hat{\beta}_0^*$

is simply the reciprocal of the sum of the weights,

$$v_\bullet^* = \left(\sum_{i=1}^{k} w_i^* \right)^{-1}, \qquad (24)$$

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

$$\hat{\beta}_0^* - t_{\alpha/2}\sqrt{v_\bullet^*} \leq \beta_0^* \leq \hat{\beta}_0^* + t_{\alpha/2}\sqrt{v_\bullet^*}, \quad (25)$$

where *t*
#

is the $_{100}\alpha$ percent point of the *t* distribution with *(k* – 1) degrees of freedom. Similarly, a two–sided test of the hypothesis that β
0

* = 0 at significance level α uses the test statistic

$$Z = = \hat{\beta}_0^* / \sqrt{}$$

and rejects if |Z | exceeds t
a/2

. Note that the same test and confidence intervals can be computed **[p. 185 ↓]** 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 ith study is modeled as

$$\theta_i = \beta_1^* x_{i1} + \cdots + \beta_p^* x_{ip} + \eta_i,\ \eta_i \sim N(0, \tau^2), \quad (26)$$

where η_i is a study–specific random effect with zero expectation and variance τ^2 (and all other terms have been defined previously).

Then, the T_i is modeled as

$$\begin{aligned} T_i &= \beta_1^* x_{i1} + \cdots + \beta_p^* x_{ip} + \eta_i + \varepsilon_i \\ &= \beta_1^* x_{i1} + \cdots + \beta_p^* x_{ip} + \xi_i, \end{aligned} \quad (27)$$

where $\xi_i = \eta_i + \#_i$ is a composite residual incorporating both study–specific random effect and sampling error. Because we assume that η_i and $\#_i$

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 (). The description of the weighted least squares estimation is facilitated by describing the model in matrix notation, and as in fixed effects models to compute the regression coefficients, we use the method of generalized least squares (see appendix).

Tests for Blocks of Regression Coefficients

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

$$H_0: \beta_1^* = \ldots = \beta_m^* = 0. \qquad (28)$$

This test arises, for example, in stepwise analyses, where it is desired to determine whether a set of *m* of the *p* predictor variables ($m \leq p$) is related to the outcome after controlling for the effects of the other predictor variables. To test this hypothesis, we compute the statistic

$$Q^* = (\hat{\beta}_1^*, \ldots, \hat{\beta}_m^*), (\boldsymbol{\Sigma}_{11}^*)^{-1} (\hat{\beta}_1^*, \ldots, \hat{\beta}_m^*)', \quad (29)$$

where (σ_{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 of 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. In this case, the test statistic *Q#* given in Equation 29 becomes the weighted sum of squares due to regression (see appendix).

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

$$Q = \sum_{i=1}^{k} (T_i - \hat{\beta}_0)^2 / v_i, \qquad (30)$$

where

$\hat{\beta}_0$

is the estimate of β_0

that would be obtained 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 a rejects the hypothesis if Q exceeds the $100(1 - \alpha)$ 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 & 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 of 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 **[p. 186 ↓]** 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

$$\hat{\tau}^2 = \begin{bmatrix} \frac{Q-(k-1)}{a} & if\ Q \geq (k-1) \\ 0 & if\ Q < (k-1) \end{bmatrix}, (31)$$

where a is given by

$$a = \sum_{j=1}^{k} w_i - \frac{\sum_{j=1}^{k} w_i^2}{\sum_{j=1}^{k} w_i}, \qquad (32)$$

and w_i = 1/v. 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

(see appendix). If the null hypothesis

$$H_0: \tau^2 = 0 \qquad (33)$$

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 a 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 12.1). 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 $Q = 24.10$, which is slightly larger than 22.36, which is the 100(1 – .05) = 95% point of the chi–square distribution with 14

– 1 = 13 degrees of freedom. Actually, a *Q* value of 24.10 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 size estimates across studies) using the distribution–free method described above and

$$\hat{\tau}^2 = (24.10 - (14 - 1)/195.38 = 0.06$$

. 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

$$\hat{\beta}_0^* = 0.55$$

with a variance of *v*

.

#. 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 $0.34 \leq \beta^*$
0

$\leq$ 0.76.This confidence interval does not include 0, so the data are incompatible with the hypothesis that β^*_0 = 0. Alternatively, the ratio

$$\hat{\beta}^*_0 / \sqrt{v^*_\bullet} = 5.5,$$

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 α = .05 significance level (2.16).

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 **[p. 187 ↓]** 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 the between–study variance component given, for example, by DerSimonian and Laird (1986). Specifically, the usual estimator of the residual variance component is given by

$$\hat{\tau}^2 = (Q_E - k + p)/c, \qquad (34)$$

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*- = 1/v
i

for each study), and *c* is defined in the appendix.

First we compute the constant *c* as *c* = 174.54 and the *Q*
E

as *Q*
E

= 15.11. Hence,

$$\hat{\tau}^2 = (15.11 - 12)/174.54 = 0.018$$

, which is nearly 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

$$\hat{\beta}_0^* = 3.22$$

for the intercept term and

$$\hat{\beta}_1^* = -0.04$$

for the effect of year. The variances of the regression estimates are 1.26 for the intercept term and 0.0003 for the year of study effect. The 95% confidence interval for

$\hat{\beta}_1^*$

is given by -0.08 $\leq \beta$
1

$^{*} \leq$ -0.004. This confidence interval does not include 0, so the data are incompatible with the hypothesis that $\beta^{*1} = 0$. Alternatively, the ratio

$$\hat{\beta}_1^* / \sqrt{\text{var}(\hat{\beta}_1^*)} = -2.3,$$

which indicates that the year effect is significantly different from zero since the absolute observed value is larger than the two–tailed critical *t* value at the $\alpha = .05$ significance level with 12 degrees of freedom (2.18). 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 CMA by Hedges et al. (2005).

Multivariate Meta–Analysis

In the previous sections, we portrayed methods for fitting general linear models to the effect sizes from a series of studies when the effect size estimates are independent. This assumption is reasonable when each study provides only one effect size estimate (e.g., a correlation coefficient). However, there are cases where studies provide information on two or more effect size estimates. In such cases, the effect size estimates are correlated, and hence the sampling errors are not independent. Appropriate analyses should take this correlation between the effect size estimates into account. In this section, we sketch analogues to the methods portrayed in previous sections when the sampling errors are not independent. These methods are essentially multi–variate generalizations of the fixed and mixed effects models given above for univariate meta–analysis. To use these methods, the joint distribution of the nonindependent 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.

In the multivariate case, assuming there are *q* effect size estimates in each study, the effect size parameter θ_{ij}

for the ith study and the jth estimate is modeled as

$$\theta_{ij} = \beta_{1j}x_{i1} + \cdots + \beta_{pj}x_{ip},\ i = 1, \ldots, k;\ j = 1, \ldots, q, \qquad (35)$$

where β_{1j}

,…,β_{pj}

are unknown regression coefficients that need to be estimated. Hence, the T–in each study is modeled as

$$T_{ij} = \beta_{1j} x_{i1} + \cdots + \beta_{pj} x_{ip} + \varepsilon_{ij}. \qquad (36)$$

To compute the regression coefficients, we use the generalized least squares, which is also **[p. 188 ↓]** the maximum likelihood estimator (see appendix). Once the regression estimates and their standard errors are computed, one can construct tests and confidence intervals for individual regression coefficients or tests for blocks of regression coefficients that are similar to those used in the univariate fixed effects models. Tests of goodness of fit of regression models are straightforward generalizations of those used in the univariate general linear model.

Example: Studies of the Effects of Coaching on the SAT

A collection of 19 studies of the effects of coaching on SAT verbal and mathematics scores was assembled by Kalaian and Raudenbush (1996). The authors 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 effect size estimates are standardized mean differences expressing the difference in SAT mathematics or verbal scores between students who received coaching and students who did not receive any coaching. These data are summarized in Table 12.2. Positive estimates indicate the benefits of coaching, while negative estimates indicate higher performance for students who did not receive coaching. Using the formulas illustrated in the appendix, we first compute the estimates of the regression coefficients as

$$\hat{\beta}_1 = -0.13$$

(the intercept for SAT verbal standardized mean differences),

$$\hat{\beta}_2 = 0.08$$

(the association between hours of coaching and SAT verbal standardized mean differences),

$$\hat{\beta}_3 = -0.29$$

(the intercept for SAT mathematics standardized mean differences), and

$$\hat{\beta}_4 = 0.13$$

(the association between hours of coaching and SAT verbal standardized mean differences). Then, we compute the standard errors of the coefficients as

$$SE(\hat{\beta}_1) = \sqrt{\sigma}_{11} = 0.22$$

,

$$SE(\hat{\beta}_2) = \sqrt{\sigma}_{22} = 0.07$$

,

$$SE(\hat{\beta}_3) = \sqrt{\sigma}_{33} = 0.22$$

and

$$SE(\beta_4) = \sqrt{\sigma}_{44} = 0.07$$

. Finally, we compute the individual test statistics for the four regression coefficients and obtain t
j

= -0.59, t
2

= 1.12, t

3

= -1.34, and t
4

= 1.91. Notice that none of the two–tailed tests is statistically significant at the .05 significance level, except *t*
4

if we assume a one–tailed test. Hence, it looks like hours of coaching is not significantly associated with SAT mathematics or verbal effect size estimates.

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. In this case, there is greater residual covariation than would be expected from sampling variability alone, which indicates systematic variation between studies. 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. The multivariate version of mixed effects models is a straightforward extension of the univariate case.

Let's assume that there are *q* effect size estimates in each study. Then, the effect size parameter θ_{ij} for the ith study is modeled as

$$\theta_{ij} = \beta^{*}_{1j} x_{i1} + \cdots + \beta^{*}_{pj} x_{ip} + \xi_{ij}, \quad i = 1, \ldots, k; j = 1, \ldots, q, \qquad (37)$$

where β_{1j},…,β^{*}_{pj} are unknown regression coefficients that need to be estimated, and ξ_{ij} are study–and effect size–specific random effects. Hence, the Tj in each study is modeled as

$$T_{ij} = \beta^{*}_{1j} x_{i1} + \cdots + \beta^{*}_{pj} x_{ip} + \xi_{ij} + \varepsilon_{ij}. \quad (38)$$

Estimation of the Regression Coefficients and the Covariance Components

The regression coefficients and the covari–ance components 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 components and then reweight to estimate the regression coefficients and their standard errors. 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 (HLM; see Thum, 1997), estimating parameters by the method of maximum likelihood. However, a simpler alternative is available since **[p. 189 ↓]** the sampling error covariance matrix is known (Kalaian & Raudenbush, 1996). In particular, it is possible to transform the within–study model so that the sampling errors are independent (see appendix). Eventually, the model that results from this procedure resembles a conventional two–level linear model with independent sampling errors at the first level. Therefore, conventional software can

be used to estimate the regression coefficients and the variance components (such as HLM).

Table 12.2 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

Multivariate Meta–Analysis Using HLM

HLM is a software package designed especially for fitting multilevel models, and it can be used to fit mixed effects models to effect size data with study–level covariates (Raudenbush, Bryk, Cheong, & Congdon, 2005; readers can refer to Chapter 31 for more information on this application of HLM). It can also be used to fit multivariate

mixed models to effect size data in meta–analysis. Table 12.3 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 two lines set the maximum number of iterations the program will run (NUMIT:1000) and the criterion for stopping iteration (STOP–VAL:0.0000010000), and the third line specifies that a linear model will be used (NONLIN: n). Lines 4 to 6 indicate the Level 1 model (LEVEL 1: MATH = VERBAL + MATH + RANDOM) and the Level 2 models (LEVEL 2: VERBAL = INTR–CPT2 + HOURS + RANDOM/and LEVEL 2: MATH = INTRCPT2 + HOURS + RANDOM/). Lines 7 and 8 indicate that no weights are used in **[p. 190 ↓]** the computations (LEVELWEIGHT:NONE). Line 9 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 is constrained to be equal to one another (CON–STRAIN:N). Line 19 specifies that the output file is named “COACHING.OUT,” line 20 specifies that the full output will be given (FULL–OUTPUT:Y), and line 21 specifies the title of the output.

The results are reported in Table 12.4. The top panel of Table 12.4 shows the regression coefficient estimates. The estimates are only slightly different from 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 12.4 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. Variance components for both math

and verbal are not significantly different from zero, indicating that there is negligible between–study variation. This indicates that a fixed effects model is appropriate.

Table 12.3 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

Conclusion

This study presented univariate and multivariate models for meta–analysis. The use of fixed and mixed effects models in univariate and multivariate cases was also demonstrated. Specialized statistical software packages such as CMA can be easily used to conduct univariate weighted least squares analyses in meta–analysis (for both 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).

Table 12.4 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	*Standard Coefficient*	*Error*	*Approximate T Ratio*	df	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

Appendix

Univariate Meta–Analysis

Fixed Effects Models

The model in Equation 18 can be written in matrix notation as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (A1)$$

where $\theta = (\theta_1, \ldots, \theta_k)'$ and $T = (T_1, \ldots, T_k)'$ denote the κ–dimensional vectors of population and sample effect sizes, respectively; $\beta = (\beta_1, \ldots, \beta_p)'$ is the p–dimensional vector of regression coefficients; # = (#

,…, #
k

)' = **T** – θ is a k–dimensional vector of residuals; and **X** is a *k* × *p* matrix

$$X = \begin{bmatrix} 1\ x_{12} \ldots x_{1p} \\ 1\ x_{22} \ldots x_{2p} \\ .\ \ .\ \ \ldots\ . \\ .\ \ .\ \ \ldots\ . \\ 1\ x_{k2} \ldots x_{kp} \end{bmatrix} \qquad \text{(A2)}$$

called the *design matrix,* which is assumed to have no linearly dependent columns. The generalized least squares estimator β, which is also the maximum likelihood estimator of β, is given by

$$\hat{\boldsymbol{\beta}} = (\mathrm{X}'\mathrm{V}^{-1}\,\mathrm{X})^{-1}\,\mathrm{X}\mathrm{V}^{-1}\mathrm{T}, \qquad \text{(A3)}$$

which has a normal distribution, with mean β and covariance matrix σ given by

$$\boldsymbol{\Sigma} = (\mathrm{X}'\mathrm{V}^{-1}\,\mathrm{X})^{-1}, \qquad \text{(A4)}$$

where **V** is a diagonal covariance matrix,

$$\mathrm{V} = \mathrm{Diag}(v_1, v_2, \ldots, v_k). \qquad \text{(A5)}$$

Testing the Significance of All Regression Coefficients

When a meta–analyst is interested in testing whether all the β
j

S are simultaneously zero, the test statistic becomes the weighted sum of squares due to regression, namely,

$$Q_R = \hat{\boldsymbol{\beta}}' \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\beta}}. \qquad (A6)$$

The test that $\beta = 0$ is simply a test of whether the weighted sum of squares due to the regression is larger than would be expected if $\beta = 0$, and the test consists of rejecting the hypothesis that $\beta = 0$

[p. 191 ↓] if *Q*
r

exceeds the 100(1 – α) percentage point of a chi–square with *p* degrees of freedom.

Mixed Effects Models

The model in Equation 27 can be written in matrix notation as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}, \quad (A7)$$

where η = (η
1

’…,η

k

), is the κ–dimensional vector of random effects, and $\xi=(\xi$
i

,…,ξ
K

) is a κ–dimensional vector of residuals of **T** about **X#*** and all other terms have been defined previously. The covariance matrix of ξ is a diagonal matrix where the *i*th diagonal element is

$$\nu_i + \hat{\tau}^2$$

. 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 compute the generalized least squares estimate of β^*—namely,

$$\hat{\beta}^{*}—$$

—as

$$\hat{\beta}^{*} = [X'(V^{*})^{-1}X]^{-1}\, X(V^{*})^{-1}T, \qquad (A8)$$

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

$$\Sigma^{*} = [X'(V^{*})^{-1}X]^{-1}, \qquad (A9)$$

where **V#** is defined as

$\mathbf{V}^*$

That is, 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.

Testing the Significance of All Regression Coefficients

When a meta–analyst is interested in testing whether all the ßjs are simultaneously zero, the test statistic becomes the weighted sum of squares due to regression, namely,

$$Q_R^* = (\hat{\boldsymbol{\beta}}^*)' \, (\boldsymbol{\Sigma}^*)^{-1} \, \hat{\boldsymbol{\beta}}^*. \qquad \text{(A11)}$$

The test that β^* = **0** is simply a test of whether the weighted sum of squares due to the regression is larger than would be expected if β^* = **0,** and the test consists of rejecting the hypothesis that β^* = **0 [p. 192 ↓]** if Q*R exceeds the 100(1 – α) percentage point of a chi–square with *p* degrees of freedom.

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

$$Q_E = \mathbf{T}'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{T}, \quad (A12)$$

where **V** = Diag(v
1

,…, v
k

). This statistics is also used to compute the residual variance component

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

where *c* is given by

$$c = \mathrm{tr}(\mathbf{V}^{-1}) - \mathrm{tr}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-2}\mathbf{X}], \quad (A13)$$

where tr**(A)** is the trace of the matrix **A.**

Multivariate Meta–Analysis

Fixed Effects

Equation 36 can be expressed in matrix notation as

$$\mathbf{T} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (A14)$$

where we denote the #q–dimensional column vectors of population and sample effect sizes by $\theta = (\theta$
1

,…,θ^{k}) and **T** = **(T'**
1

,…,T'
k

), respectively, where θ = (θ
1

,…,θ^{k}) and T
i

= (T
il

,…,T
iq

);

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

is a #q×pq design matrix, where ***x***
i
= (x
il

,…,x
ip

); **I,**
q

is a *q* × *q* identity matrix; ⊗ is the Kronecker operator; β=(β
1l

,…,β
pl

,β
2l

,.β
pq

) is a pq column vector of regression coefficients that need to be estimated; and # = (#
1

,…,#
kq

)' = **T**– θ is a kq–dimensional column vector of residuals. Each **T**
i

is assumed to have a q–variate normal distribution (since there are *q* effect size estimates in each study) about the corresponding θ
i

with known *q* × *q* covariance matrix σ
i

. Although there is no need for all studies to have the same number of effect sizes, we make that assumption here to simplify notation.

[p. 193 ↓] The vector of residuals # = **T** – θ follows a *kq*-variate normal with mean zero and known *kq* × *kq* block–diagonal covariance matrix **V** given by

$$\begin{aligned} \mathbf{V} &= \text{Diag}\,(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_k) \\ &= \mathbf{I}_k \otimes \boldsymbol{\Sigma}_i \; i = 1, \ldots, k \end{aligned} \qquad \text{(A15)}$$

where Σ

i

is a known *q* × *q* covariance matrix for study *i.* We can hence use the method of generalized least squares to obtain an estimate of the regression coefficients vector β. This is essentially the approach employed by Raudenbush, Becker, and Kalaian (1988); Gleser and Olkin (1994); and Berkey, Anderson, and Hoaglin (1996). Specifically, the generalized least squares estimator

$\hat{\beta}$

, which is also the maximum likelihood estimator of

$\hat{\beta}$

, with covariance matrix **V,** is given by

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}T, \qquad (A16)$$

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

$$\Sigma = (X'V^{-1}X)^{-1}. \qquad (A17)$$

Mixed Effects

Equation 38 can be expressed in matrix notation as

$$T = \theta + \varepsilon = X\beta^{*} + I\Xi + \varepsilon, \qquad (A18)$$

where **I** is a kq–dimensional identity matrix,

Ξ

is a kq–dimensional vector of the between–study random effects, and all other terms have been defined previously. The vector β^* of the between–study random effects follows a *q*-variate normal with mean zero and *q* × *q* covariance matrix Ω

The regression coefficient vector β^* and the covariance component matrix Ω 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 Ω and then reweight to estimate the regression coefficient vector β^* and its covariance matrix Σ^*. Alternatively, one could orthogonalize the error terms. To achieve this, one can perform the Cholesky factorizatio on each sampling error covariance matrix in each study so that

$$\boldsymbol{\Sigma}_i^* = \mathbf{F}_i \mathbf{F}_i', \qquad (A19)$$

where F
i

is a known matrix (since Σ^{*i} is a known matrix) and is the lower triangular (square root)

matrix of the Cholesky decomposition. The within–study model is then transformed to

$$\mathbf{F}_i^{-1}\,\mathbf{T}_i = \mathbf{F}_i^{-1}\,\boldsymbol{\theta}_i + \mathbf{F}_i^{-1}\,\boldsymbol{\varepsilon}_i, \qquad (A20)$$

where the transformed effect size vector Z
i

is given by

$$\mathbf{Z}_i = \mathbf{F}_i^{-1}\,\mathbf{T}_i \qquad (A21)$$

and has a sampling error vector

$$\tilde{\varepsilon}_i = \mathbf{F}_i^{-1}\,\boldsymbol{\varepsilon}_i, \qquad (A22)$$

which has covariance matrix **I,** a *q*
i

× *q*
i

identity matrix. Thus, one might write the model as

$$\mathbf{Z}_i = \mathbf{F}_i^{-1}\,\boldsymbol{\theta}_i + \tilde{\varepsilon}_i, \qquad (A23)$$

where the transformed effect size estimates **Z**
i

are now independent with a constant variance, and the effect size parameter vector θ
;

is the same as in the original model. Thus, the within–study model 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 β^* and Ω by the method of maximum likelihood such as HLM (Raudenbush et al., 2005).

Note

1. Comprehensive Meta–Analysis offers about 100 different formats for entering data and is especially designed to cover various methods for meta–analytic data (see http://www.meta–analysis.com).

References

Begg, C. B. (1994). Publication bias . In H. Cooper, ed. & L. V. Hedges (Eds.), The handbook of research synthesis (pp. p. 399-410). New York: Russell Sage Foundation.

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

Cohn, L. D. and Becker, B. J. Title: How metaanalysis increases statistical power Psychological Methods, vol. 8, p. 243-253(2003).

Cooper, H. (1989). Integrating research (2nd ed.). Newbury Park, CA: Sage.

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

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

Duval, S. and Tweedie, R. A nonparametric trim and fill method of accounting for publication bias in meta-analysis Journal of the American Statistical Association, vol. 95, p. 89-98(2000).

Fleiss, J. L. (1994). Measures of effect size for categorical data . In H. Cooper, ed. & L. V. Hedges (Eds.), The handbook of research synthesis (pp. p. 245-260). New York: Russell Sage Foundation.

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

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

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

Hedges, L. V. (1994). Fixed effects models . In H. Cooper, ed. & L. V. Hedges (Eds.), The handbook of research synthesis (pp. p. 285-299). New York: Russell Sage Foundation.

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

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

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

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

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

Kalaian, H. and Raudenbush, S. W. A multivariate mixed linear model for metaanalysis Psychological Methods, vol. 1, p. 227-235(1996).

Konstantopoulos, S. (2005, April). Three-level models in meta-analysis . Paper presented at the annual conference of the American Educational Association, Montreal, Canada.

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

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

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

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

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. The “file-drawer problem” and tolerance for null results Psychological Bulletin, vol. 86, p. 638-641(1979).

Rosenthal, R. (1994). Parametric measures of effect size . In H. Cooper, ed. & L. V. Hedges (Eds.), The handbook of research synthesis (pp. p. 231-244). New York: Russell Sage Foundation.

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

Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis: Prevention, assessment, and adjustments . Hoboken, NJ: John Wiley.

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

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

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size . In H. Cooper, ed. & L. V. Hedges (Eds.), The handbook of research synthesis (pp. p. 261-281). New York: Russell Sage Foundation.

Smith, M. I. and Glass, G. V. Meta-analysis of psychotherapy outcome studies American Psychologist, vol. 32, p. 752-760(1977).

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

White, H. D. (1994). Scientific communication and literature retrieval . In H. Cooper, ed. & L. V. Hedges (Eds.), The handbook of research synthesis (pp. p. 41-55). New York: Russell Sage Foundation.

Wortman, P. M. (1994). Judging research quality . In H. Cooper, ed. & L. V. Hedges (Eds.), The handbook of research synthesis (pp. p. 97-110). New York: Russell Sage Foundation.

10.4135/9781412995627.d15