

The Multilevel Approach to Meta-Analysis

Meta-analysis is a systematic approach towards summarizing the findings of a collection of independently conducted studies on a specific research problem. In meta-analysis, statistical analyses are carried out on the published results of empirical studies on a specific research question. This chapter shows that multilevel regression models are attractive for analyzing meta-analytic data.

10.1 META-ANALYSIS AND MULTILEVEL MODELING

Meta-analysis is a systematic approach towards the synthesis of a large number of results from empirical studies (cf. Glass, 1976, Lipsey & Wilson, 2001). The goal is to summarize the findings of a collection of independently conducted studies on a specific research problem. For instance, the research question might be: ‘What is the effect of social skills training on socially anxious children?’ In a meta-analysis, one would collect reports of experiments concerning this question, explicitly code the reported outcomes, and integrate the outcomes statistically into a combined ‘super outcome’. Often the focus is not so much on integrating or summarizing the outcomes, but on more detailed questions such as: ‘What is the effect of different durations for the training sessions?’ or ‘Are there differences between different training methods?’ These questions address the generalizability of the research findings. In these cases, the meta-analyst not only codes the study outcomes, but also codes study characteristics. These study characteristics are potential explanatory variables to explain differences in the study outcomes. Meta-analysis is not just the collection of statistical methods used to achieve integration. It is the application of systematic scientific strategies to the literature review (Cornell & Mulrow, 1999; Light & Pillemer, 1984). For a brief introduction to general meta-analysis I refer to Cornell and Mulrow (1999), and Lipsey and Wilson (2001). A thorough and complete treatment of methodological and statistical issues in meta-analysis, including a chapter on using multilevel regression methods (Raudenbush, 1994) can be found in Cooper and Hedges (1994) and in Sutton et al. (2000).

The core of meta-analysis is that statistical analyses are carried out on the published results of a collection of empirical studies on a specific research question. One approach is to combine the p -values of all the collected studies into one combined p -value. This is a simple matter, but does not provide much information. A very general model for meta-analysis is the random-effects model (Hedges & Olkin, 1985, p. 198). In this model, the focus is not on establishing the statistical significance of a combined outcome, but on analyzing the variation of the effect sizes found in the different studies. The random-effects model for meta-analysis assumes that study outcomes vary across studies, not only because of random sampling effects, but also because there are real differences between the studies. For instance, study outcomes may vary because the different studies employ different sampling methods, use different experimental manipulations, or measure the outcome with different instruments. The random-effects model is used to decompose the variance of the study outcomes into two components: one component that is the result of sampling variation, and a second component that reflects real differences between the studies. Hedges and Olkin (1985) and Lipsey and Wilson (2001) describe procedures that can be used to decompose the total variance of the study outcomes into random sampling variance and systematic between-studies variance, and procedures to test the significance of the between studies variance. If the between studies variance is large and significant, the study outcomes are regarded as *heterogeneous*. This means that the studies do not all provide the same outcome. One procedure to investigate the differences between the studies is to form clusters of studies, which differ in their outcomes between the clusters, but which are homogeneous within the clusters. The next goal is to identify study characteristics that explain differences between the clusters. Variables that affect the study outcomes are in fact moderator variables: variables that interact with the independent variable to produce variation in the study outcomes.

Meta-analysis can be viewed as a special case of multilevel analysis. We have a hierarchical data set, with subjects within studies at the first level, and studies at the second level. If the raw data of all the studies would be available, we could carry out a standard multilevel analysis, predicting the outcome variable using the available individual and study level explanatory variables. In our example on the effect of social skills training of children, we would have one outcome variable, for instance the result on a test measuring social skill, and one explanatory variable which is a dummy variable that indicates whether the subject is in the experimental or the control group. On the individual level, we have a linear regression model that relates the outcome to the experimental/ control-group variable. The general multilevel regression model assumes that each study has its own regression model. If we have access to all the original data, standard multilevel analysis can be used to estimate the mean and variance of the regression coefficients across the studies. If the variance of the regression slopes of the experimental/control-group variable is large and significant, we have heterogeneous results. In that case, we can use the available study characteristics as explanatory variables at the second (study) level to predict the differences of the regression coefficients.

These analyses can be carried out using standard multilevel regression methods and using standard multilevel software. However, in meta-analysis we usually do *not* have access to the original raw data. Instead, we have the

published results in the form of p -values, means, standard deviations or correlation coefficients. Classical meta-analysis has developed a large variety of methods to integrate these statistics into one overall outcome, and to test whether the outcomes should be regarded as homogeneous or heterogeneous. Hedges and Olkin (1985) discuss the statistical models on which these methods are based. Hedges and Olkin describe a weighted regression model that can be used to model the effect of study characteristics on the outcomes, and Lipsey and Wilson (2001) show how conventional software for weighted regression analysis can be used to analyze meta-analytic data. Hunter and Schmidt (2004) discuss different approaches, but also use random effects models that resemble multilevel analysis.

Even without access to the raw data, it is often possible to carry out a multilevel meta-analysis on the summary statistics that are the available data for the meta-analysis. Raudenbush and Bryk (Raudenbush & Bryk, 1985; Raudenbush & Bryk, 2002) view the random-effects model for meta-analysis as a special case of the multilevel regression model. The analysis is performed on sufficient statistics instead of raw data, and as a result, some specific restrictions must be imposed on the model. In multilevel meta-analysis, it is simple to include study characteristics as explanatory variables in the model. If we have hypotheses about study characteristics that influence the outcomes, we can code these and include them on a priori grounds in the analysis. Alternatively, after we have concluded that the study outcomes are heterogeneous, we can explore the available study variables in an attempt to explain the heterogeneity.

The major advantage of using multilevel analysis instead of classical meta-analysis methods is flexibility (Hox & de Leeuw, 2002). Using a multilevel framework, it is easy to add further levels to the model, for example to accommodate multiple outcome variables. Estimation can be done using Maximum Likelihood methods, and a range of estimation and testing methods are available (cf. Chapters Three and XXX). However, not all software can be used for multilevel meta-analysis; the main requirement is that it is possible to impose constraints on the random part of the model.

10.2 THE VARIANCE-KNOWN MODEL

In a typical meta-analysis, the collection of studies found in the literature employs different instruments and use different statistical tests. To make the outcomes comparable, the study results must be transformed into a standardized measure of the effect size, such as a correlation coefficient or the standardized difference between two means. For instance, if we perform a meta-analysis on studies that compare an experimental group to a control group, an appropriate measure for the effect size is the standardized difference between two means g , which is given by $g = (\bar{Y}_E - \bar{Y}_C) / s$. The standard deviation s is either the standard deviation in the control group, or the pooled standard deviation for both the experimental and control group. Since the standardized difference g has a small upwards bias, it is often transformed to the unbiased effect size indicator $d = (1 - 3/(4N - 9))g$, where N is the total sample size for the study. This correction is most appropriate when N is less than 20, with larger sample sizes the bias correction is often negligible (Hedges & Olkin, 1985).

The general model for the study outcomes, ignoring possible study-level explanatory variables, is given by

$$d_j = \delta_j + e_j \quad (10.1)$$

In equation 10.1, d_j is the outcome of study j ($j=1, \dots, J$), δ_j is the corresponding population value, and e_j is the sampling error for this specific study. It is assumed that the e_j have a normal distribution with a known variance σ_j^2 . If the sample sizes of the individual studies are not too small, for instance between 20 (Hedges & Olkin, 1985, p. 175) to 30 (Raudenbush & Bryk, 2002, p. 207), it is reasonable to assume that the sampling distribution of the outcomes is normal, and that the known variance can be estimated from the data with sufficient accuracy. The assumption of underlying normality is not unique for multilevel meta-analysis; most classical meta-analysis methods also assume normality (cf. Hedges & Olkin, 1985). The variance of the sampling distribution of the outcome measures is assumed known from statistical theory.

Table 10.1 Some effect measures, their transformation and sampling variance

Measure	Estimator	Transformation	Sampling variance
mean	\bar{x}	-	s^2/n
diff. 2 means	$g = (\bar{y}_E - \bar{y}_C) / s$	$d = (1 - 3/(4N - 9))g$	$(n_E + n_C) / (n_E n_C) + d^2 / (2(n_E + n_C))$
stand. dev.	s	$s^* = \text{LN}(s) + 1/(2df)$	$1/(2df)$
correlation	r	$z = 0.5 \text{LN}((1+r)/(1-r))$	$1/(n-3)$
proportion	p	$\text{logit} = \text{LN}(p/(1-p))$	$1/(np(1-p))$

diff. 2 prop.	$d \approx Z_{p_1} - Z_{p_2}$	-	$\frac{2\pi p_1(1-p_1)e^{Z_{p_1}^2}}{n_1} +$ $\frac{2\pi p_2(1-p_2)e^{Z_{p_2}^2}}{n_2}$
diff. 2 prop.	Log Odds Ratio	-	$\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$ a, b, c, d are cell freq.
reliability coefficient alpha	Cronbach's α	$Z = \text{LN}(1- \alpha)$	$\frac{k}{2(k-1)(n-2)}$ $k = \# \text{ of items}$

To obtain a good approximation to a normal sampling distribution, and to determine the known variance, a transformation of the original effect size statistic is often needed. For instance, since the sampling distribution of a standard deviation is only approximately normal, it should not be used with small samples. The transformation $s^* = \text{LN}(s) + 1/(2df)$ of the standard deviation improves the normal approximation. The usual transformation for the correlation coefficient r is the familiar Fisher-Z transformation, and for the proportion it is the logit. Note that, if we need to perform a meta-analysis on logits, the procedures outlined in Chapter Six are generally more accurate. Usually, after a confidence interval is constructed for the transformed variable, the end-points are translated back to the scale of the original estimator. Table 10.1 lists some common effect size measures, the usual transformation if one is needed, and the sampling variance (of the transformed outcome if applicable) (Bonett, 2002; Lipsey & Wilson, 2001; Raudenbush & Bryk, 2002; Rosenthal, 1994).

Equation 10.1 shows that the effect sizes δ_j are assumed to vary across the studies. The variance of the δ_j is explained by the regression model

$$\delta_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \dots + \gamma_p Z_{pj} + u_j \quad (10.2)$$

where $Z_1 \dots Z_p$ are study characteristics, $\gamma_1, \dots, \gamma_p$ are the regression coefficients, and u_j is the residual error term, which is assumed to have a normal distribution with variance σ_u^2 . By substituting 10.2 into 10.1 we obtain the complete model

$$d_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \dots + \gamma_p Z_{pj} + u_j + e_j \quad (10.3)$$

If there are no explanatory variables, the model reduces to

$$d_j = \gamma_0 + u_j + e_j \quad (10.4)$$

Model 10.4, which is the 'intercept only' or 'empty' model, is equivalent to the random-effects model for meta-analysis described by Hedges and Olkin (1985).

In model 10.4, the intercept γ_0 is the estimate for the mean outcome across all studies. The variance of the outcomes across studies, σ_u^2 , indicates how much these outcomes vary across studies. Thus, testing if the study outcomes are homogeneous is equivalent to testing the null-hypothesis that the variance of the residual errors u_j , indicated by σ_u^2 , is equal to zero. If the test of σ_u^2 is significant, the study outcomes are considered heterogeneous. The proportion of systematic between-study variance can be estimated by the intraclass correlation $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$.

The general model 10.3 includes study characteristics Z_{pj} to explain differences in the studies' outcomes. In model 10.3, σ_u^2 is the residual between-study variance after the explanatory variables are included in the model. A statistical test on σ_u^2 now tests whether the explanatory variables in the model explain all the variation in the studies' outcomes, or if there still is unexplained systematic variance left in the outcomes. The difference between the between-studies variance σ_u^2 in the empty model and in the model that includes the explanatory variables Z_{pj} , can be interpreted as the amount of variance explained by the explanatory variables, that is, by the study characteristics.

The multilevel meta-analysis model given by equation 10.3 is equal to the general weighted regression model for random effects described by Hedges and Olkin (1985, chapter 9). When the study level variance is not significant, it can be removed from the model:

$$d_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \dots + \gamma_p Z_{pj} + e_j. \quad (10.5)$$

Compared to model 10.3, model 10.5 lacks the study-level residual error term u_j . The result is called the fixed effect model, which assumes that all studies are homogeneous and all estimate the same underlying population parameter δ . Thus, the fixed effect model described by Hedges and Olkin (1985, chapter 8) is a special case of the random effects weighted regression or the multilevel meta-analysis model. Omitting the study-level residual error term u_j implies that there is no variation in the effect sizes across all studies, or that the explanatory variables in the model explain all the variance among the studies. Thus, if the residual between-study variance is zero, a fixed-effect model is appropriate

(Hedges & Vevea, 1998). However, this assumption is not very realistic. For instance, Hunter and Schmidt (2004) argue that the between-studies heterogeneity is partly produced by some unavoidable artifacts encountered in meta-analysis. Examples of such artifacts are the (usually untestable) assumption of a normal distribution for the sampling errors e_j , the correctness of statistical assumptions made in the original analyses, differences in reliability of instruments used in different studies, coder unreliability, et cetera. It is unlikely that the available study level variables cover all these artifacts. Generally, the amount of detail in the input for the meta-analysis, the research reports, papers and articles, is not enough to code all these study characteristics for all of the studies. Therefore, heterogeneous results are to be expected (cf. Engels, Schmidt, Terrin, Olkin & Lau, 2000). Since heterogeneous results are common, Hunter and Schmidt recommend as a rule of thumb that the study-level variance should be larger than 25% of all variance to merit closer inspection; study level variance smaller than 25% is likely to be the result of methodological differences between the studies. However, simulations have shown that this '25% rule' is very inaccurate and therefore not recommended (Schulze, 2008).

Since there is as a rule unexplained variance in a meta-analysis, random-effects models should be preferred over fixed-effect models. Lipsey and Wilson (2001) describe a Weighted Least Squares regression procedure for estimating the model parameters, which can be applied using standard statistical software for weighted regression. Just like multilevel meta-analysis this is a powerful approach, because one can include explanatory variables in the model. However, in the weighted regression approach the investigators must supply an estimate of the between-study variance. This variance is estimated before the weighted regression analysis, and the estimated value is then plugged into the weighted regression analysis (Lipsey & Wilson, 2001). Multilevel analysis programs estimate this variance component, typically using iterative Maximum Likelihood estimation, which in general is more precise and efficient. In practice, both approaches usually produce very similar parameter estimates. The multilevel approach has the additional advantage that it offers more flexibility, for example, by using a three-level model for multivariate outcomes. If fixed effect models are used in the presence of significant between-study variance, the resulting confidence intervals are biased and much too small (Villar, Mackey, Carroli & Donner, 2001; Brockwell & Gordon, 2001). If random effects models are used, the standard errors are larger, and the estimate of the average effect may be different, depending on the relation between effect sizes and sample sizes in the primary studies (cf. Villar, Mackey, Carroli & Donner, 2001).

10.3 EXAMPLE AND COMPARISON WITH CLASSICAL META-ANALYSIS

In this section we analyze an example data set using classical meta-analysis methods as implemented in the Meta Analysis macros written by David Wilson (Lipsey & Wilson, 2001, appendix D). These macros are based on methods and procedures described by Hedges and Olkin (1985). The (simulated) data set consists of 20 studies that compare an experimental group and a control group.

Table 10.2 Example meta-analytic data from 20 studies

study	weeks	g	d	$\text{var}(d)$	p	n_{exp}	n_{con}	r_{ii}
1	3	-.268	-.264	.086	.810	23	24	.90
2	1	-.235	-.230	.106	.756	18	20	.75
3	2	.168	.166	.055	.243	33	41	.75
4	4	.176	.173	.084	.279	26	22	.90
5	3	.228	.225	.071	.204	29	28	.75
6	6	.295	.291	.078	.155	30	23	.75
7	7	.312	.309	.051	.093	37	43	.90
8	9	.442	.435	.093	.085	35	16	.90
9	3	.448	.476	.149	.116	22	10	.75
10	6	.628	.617	.095	.030	18	28	.75
11	6	.660	.651	.110	.032	44	12	.75
12	7	.725	.718	.054	.003	41	38	.90
13	9	.751	.740	.081	.009	22	33	.75
14	5	.756	.745	.084	.009	25	26	.90
15	6	.768	.758	.087	.010	42	17	.90
16	5	.938	.922	.103	.005	17	39	.90
17	5	.955	.938	.113	.006	14	31	.75
18	7	.976	.962	.083	.002	28	26	.90
19	9	1.541	1.522	.100	.0001	50	16	.90
20	9	1.877	1.844	.141	.00005	31	14	.75

Let us return to our example on the effect of social skills training on socially anxious children. We collect reports of experiments concerning this question. If we compare the means of an experimental and a control group, an appropriate outcome measure is the standardized difference between the experimental and the control group, originally proposed by

Glass (1976) and defined by Hedges and Olkin as $g = (\bar{Y}_E - \bar{Y}_C) / s$, where s is the pooled standard deviation of the two groups. Because g is not an unbiased estimator of the population effect $\delta = (\mu_E - \mu_C) / \sigma$, Hedges and Olkin prefer a corrected effect measure d : $d = (1 - 3 / (4(N - 9)))g$. The sampling variance of the effect estimator d is $(n_E + n_C) / (n_E n_C) + d^2 / (2(n_E + n_C))$ (Hedges & Olkin, 1985, p. 86).

Table 10.2 is a summary of the outcomes from a collection of 20 studies. The studies are presented in increasing order of their effect sizes (g , d). Table 10.2 presents both g and d for all 20 studies, with some study characteristics. The difference between g and d is very small in most cases, where study sample sizes are larger than about 20. Table 10.2 also presents the sampling variance of the effect sizes d ($\text{var}(d)$), the one-sided p -value of the t -test for the difference of the two means (p), the number of cases in the experimental (n_{exp}) and control group (n_{con}), and the reliability (r_{ii}) of the outcome measure used in the study. The example data set contains several study level explanatory variables. A theoretically motivated explanatory variable is the duration in number of weeks of the experimental intervention. It is plausible to assume that longer interventions lead to a larger effect. In addition we have the reliability of the outcome measure (r_{ii}), and the size of the experimental and control group.

10.3.1 Classical Meta-Analysis

Classical meta-analysis includes a variety of approaches that complement each other. For instance, several different formulas are available for combining p -values. A classic procedure is the so-called Stouffer method (Rosenthal, 1991). In the Stouffer method, each individual (one-sided) p is converted to the corresponding standard normal Z -score. The Z -scores are then combined using $Z = (\sum Z_j) / \sqrt{k}$, where Z_j is the Z -value of study j , and k is the number of studies. For our example data, the Stouffer method gives a combined Z of 7.73, which is highly significant ($p < 0.0001$).

The combined p -value gives us evidence that an effect exists, but no information on the size of the experimental effect. The next step in classical meta-analysis is to combine the effect sizes of the studies into one overall effect size, and to establish the significance or a confidence interval for the combined effect. Considering the possibility that the effects may differ across the studies, the random-effects model is preferred to combine the studies.

In classical meta-analysis, the fixed effect model is used first to combine the effect sizes. It is clear that larger studies include less sampling error, and therefore deserve a larger weight in combining the effect sizes. Hedges and Olkin (1985) prove that the optimal weight is not the sample size, but the precision, which is equal to the inverse of the sampling variance. The sample size and the inverse variance weight are obviously highly related. Hence, the fixed effect model weights each study outcome with the inverse variance of the effect size: $w_j = 1 / \text{var}(d_j)$. The combined effect size is simply the weighted mean of the effect sizes. The standard error of the combined effect size is calculated as the square root of the sum of the inverse variance weights:

$$SE_{\bar{d}} = \sqrt{\frac{1}{\sum w_j}} \quad (10.6)$$

The test statistic to test for homogeneity of study outcomes is:

$$Q = \sum w_j (d_j - \bar{d})^2 \quad (10.7)$$

which has a chi-square distribution with $J-1$ degrees of freedom. If the chi-square is significant, we reject the null-hypothesis of homogeneity and conclude that the studies are heterogeneous; there is significant study level variation. In classical meta-analysis, the study level variance is estimated by a method of moments estimator given by

$$\sigma_u^2 = \frac{Q - (J - 1)}{\sum w_j - \left(\frac{\sum w_j^2}{\sum w_j} \right)} \quad (10.8)$$

The random effects model follows the same procedures, but recalculates the weights by plugging in the estimate of the study-level variance:

$$w_j^* = \frac{1}{\text{var}(d_j) + \sigma_u^2} \quad (10.9)$$

The random effects model adds the between study level variance to the known variances when calculating the inverse variance weight. Subsequently, the same methods are used to estimate the combined effect size and its standard error.

A meta-analysis of the effect sizes in Table 10.2, using the random-effects model and the methods described earlier (using the macro MEANES), estimates the overall effect as $\delta = 0.580$, with a standard error of 0.106. Using this information, we can carry out a null-hypothesis test by computing $Z = d / SE(d) = 0.58 / 0.106 = 5.47$ ($p < 0.0001$). The 95% confidence interval for the overall effect size is $0.37 < \delta < 0.79$. The usual significance test of the between-study variance

used in meta-analysis is a chi-square test on the residuals, which for our example data leads to $\chi^2=49.59$, ($df=19$, $p < 0.001$). This test is equivalent to the chi-square residuals test described by Raudenbush & Bryk (2002) and implemented in HLM. As the result is clearly significant, we have heterogeneous outcomes. This means that the overall effect 0.58 is not the estimate of a fixed population value, but an average of the distribution of effects in the population. The Z-value of 5.47 computed using the random-effects model is not the same as the Z-value of 7.73 computed using the Stouffer method. This difference is most likely due to a difference in power between these methods (Becker, 1994), since the random-effects meta-analysis produces a standard error which can be used to establish a confidence interval, we will use the results from the meta-analysis.

The parameter variance σ_u^2 is estimated as 0.14, and the proportion of systematic variance is estimated as 0.65 (estimated as σ_u^2 divided by the weighted observed variance). This is much larger than the 0.25 that Hunter and Schmidt (2004) consider a lower limit for examining differences between studies. The conclusion is that the between-study variance is not only significant, but also large enough to merit further analysis using the study characteristics at our disposal. The usual follow-up in classical meta-analysis is to use weighted regression to analyze differences between study outcomes. When the random effects model is used, the same variance estimate described earlier is plugged into the calculation of the weight, and then weighted regression methods are used to estimate regression weights for the study level variables. Instead of using the plug-in estimate, iterative Maximum Likelihood methods are also available, but they are less commonly used (cf. Lipsey & Wilson, 2001, p119). Using the method of moments estimator, the regression coefficient of the variable weeks is estimated as 0.14, with a standard error of 0.034 and an associated p -value of 0.0000. So the hypothesis that the intervention effect is larger with longer durations of the intervention is sustained. The result of the homogeneity test, conditional on the predictor weeks, is $Q=18.34$ ($df=18$, $p=0.43$). There is no evidence for study level heterogeneity once the differences in duration are accounted for. The residual variance is σ_u^2 is estimated as 0.04. The explained variance can be estimated as $(0.14-0.04)/0.14=0.71$. Given that the residual variance is not significant, one could decide to consider a fixed model one the variable weeks is included. However, the chi-square test for the between-study variance has a low power unless the number of studies is large (at least 50), so it is recommended to keep the between-studies variance in the model (cf. Huedo-Medina, Sánchez-Meca, Marín-Martínez & Botella, 2006).

10.3.2 Multilevel Meta-Analysis

A multilevel meta-analysis of the 20 studies using the empty ‘intercept only’ model produces virtually the same results as the classical meta-analysis. Since in meta-analysis we are strongly interested in the size of the between-study variance component, Restricted Maximum Likelihood (RML) estimation is the best approach.¹ Using RML, the intercept, which in the absence of other explanatory variables is the overall outcome, is estimated as $\gamma_0=0.57$, with a standard error of 0.11 ($Z=5.12$, $p<0.001$). The parameter variance σ_u^2 is estimated as 0.15 (s.e.=0.111, $Z=1.99$, $p=0.02$). As the Wald test is inaccurate for testing variances (cf. Chapter Three), the variance is also tested with the deviance difference test. This produces a chi-square of 10.61 ($df=1$, halved $p<0.001$). The proportion of systematic variance is 0.71, which is much larger than 0.25, the lower limit for examining differences between studies (Hunter & Schmidt, 2004). The differences between these results and the results computed using the classical approach to meta-analysis are small, indicating that the classical approach is quite accurate when the goal of the meta-analysis is to synthesize the results of a set of studies.

When we include the duration of the experimental intervention as an explanatory variable in the regression model, we have:

$$d_j = \gamma_0 + \gamma_1 DURATION_{ij} + u_j + e_j \quad (10.10)$$

The results of the multilevel meta-analysis are summarized in Table 10.3, which presents the results for both the empty (null) model and the model that includes duration and the results obtained by the classical (random-effects) meta-analysis.

The results are very close. It should be noted that the chi-square from the Method of Moments analysis is not straightforward: to obtain correct estimates and standard errors for the regression parameters in the second column we need to use a random effects model with the plug-in variance estimate, to obtain the correct chi-square for the residual variance we must use the fixed effect model. The multilevel analysis, using the built-in meta-analysis option in HLM, directly produces the chi-square residuals test. (Different choices and variations in software implementation are discussed in an appendix to this chapter).

Table 10.3 Results random-effects method of moments and multilevel analyses

Analysis:	Method of	Method of	Multilevel	Multilevel
	Moments	Moments	REML	REML
Delta/Intercept	.58 (.11)	-.22 (.21)	.58 (.11)	-.23 (.21)
Duration		.14 (.03)		.14 (.04)

¹ If RML is used, it is not possible to test the effect of moderator variables using the deviance test. In practice, when the difference between FML and RML estimation is small, it may be advantageous to use FML rather than RML. If the differences are appreciable, RML is recommended.

σ_u^2	.14	.04	.15 (.08)	.04 (04)
χ^2 deviance test & p -value	n/a	n/a	$\chi^2=10.6$ $p<.001$	$\chi^2=1.04$ $P=.16$
χ^2 residuals test & p -value	$\chi^2=49.6$ $p < 0.001$	$\chi^2=26.4$ $P=.09$	$\chi^2=49.7$ $p < 0.001$	$\chi^2=26.5$ $p = 0.09$

After including duration as explanatory variable in the model, the residual between-study variance is much smaller, and no longer significant. The regression coefficient for the duration is 0.14 ($p < 0.001$), which means that for each additional week the expected gain in study outcome is 0.14. The intercept in this model is -0.23 , with a standard error of 0.21 ($p = 0.27$). The intercept is not significant, which is logical, because it refers to the expected outcome of a hypothetical experiment with duration of zero weeks. If we center the duration variable by subtracting its overall mean, the intercept does not change from one model to the next, and reflects the expected outcome of the average study. The residual variance in the last model is 0.04, which is not significant. If we compare this with the parameter variance of 0.14 in the empty model, we conclude that 73% of the between-studies variance can be explained by including ‘duration’ as the explanatory variable in the model.

In the multilevel analyses reported in Table 10.3, RML estimation is used, and the residual between-studies variance is tested for significance twice, once using the deviance difference test, and once using the chi-square test proposed by Raudenbush and Bryk (2002). (The deviance test is not available in the method of moments.) As explained in more detail in Chapter Three, there are two reasons to choose for RML estimation and *not* using the Wald test on the variance. Firstly, in standard applications of multilevel analysis, the variances are often viewed as nuisance parameters. It is important to include them in the model, but their specific value is not very important, because they are not interpreted. In meta-analysis, the question whether all the studies report essentially the same outcome is an essential research question. The answer to this question depends on the size and on the decision about the significance of the between-studies variance. Therefore, it is very important to have a good estimate of the between-studies variance and its significance. For this reason, Restricted Maximum Likelihood (RML) estimation is used instead of Full Maximum Likelihood (FML). Generally, FML and RML estimation lead to very similar variance estimates, but if they do not, using RML provides better estimates (Browne, 1998). Secondly, the asymptotic Wald test on the variance computes the test statistic Z by dividing the variance estimate by its standard error. This assumes a normal sampling distribution for the variance. This assumption is not justified, because variances are known to have a chi-square sampling distribution. Compared to other tests, the Wald test of the variance has a much lower power (Berkhof & Snijders, 2001), and in general the deviance difference test is preferred (Berkhof & Snijders, 2001; LaHuis & Ferguson, 2007). The difference between the deviance difference test and the residuals chi-square test is small, unless the group sample sizes are small. A practical reason for reporting the chi-square residuals test for the variance in a meta-analysis is that the residuals chi-square test proposed by Raudenbush and Bryk (2002) follows the same logic as the chi-square test on the residuals in classic meta-analysis, which facilitates comparison.

Since the study outcome depends in part on the duration of the experiment, reporting an overall outcome for the twenty studies does not convey all the relevant information. We could report the expected outcome for different duration, or calculate which duration is minimally needed to obtain a significant outcome. This can be accomplished by centering the explanatory variable on different values. For instance, if we center the duration around two weeks, the intercept can be interpreted as the expected outcome at two weeks. Some multilevel analysis programs can produce predicted values with their expected error variance, which is also useful to describe the expected outcome for experiments with a different duration.

10.4 CORRECTING FOR ARTIFACTS

Hunter and Schmidt (1994, 2004) encourage correcting study-outcomes for a variety of artifacts. It is common to correct the outcome for the attenuation that results from unreliability of the measure used. The correction simply divides the outcome measure by the square root of the reliability, for instance $d^* = d / \sqrt{r_{ii}}$, after which the analysis is carried out as usual. This is the same correction as the classical correction for attenuation of the correlation coefficient in psychometric theory (cf. Nunnally & Bernstein, 1994). Hunter and Schmidt (2004) describe many more corrections. All these corrections share major methodological and statistical problems. One problem is that the majority of corrections always result in larger effect sizes. For instance, if the studies use instruments with a low reliability, the corrected effect size is much larger than the original effect size. If the reported reliability is incorrect, so will be the correction. Because the large effects have in fact not been observed, routinely carrying out such corrections is controversial. For that reason, Schwarzer (1989, p56) advises to always report the original values in addition to the corrected results. A second problem with all these corrections is that they influence the standard error of the outcome-measure. Lipsey and Wilson (2001) present proper standard errors for some corrections. However, if the values used to correct outcomes are themselves subject to sampling error, the sampling variance of the outcome measure becomes still larger. Especially if many corrections are performed, their cumulative effect on the bias and accuracy of the outcome-measures is totally unclear.

A different approach to correcting artifacts is to include them as covariates in the multilevel regression analysis. For reliability of the outcome measure, this is still not optimal, because the proper correction is a nonlinear multiplicative model (cf. Nunnally & Bernstein, 1994), and regression analysis is linear and additive. However, if the

range reliabilities are not extremely low (Nunnally and Bernstein suggest as a rule of thumb that the reliability of a 'good' measure should be larger than 0.70), a linear additive model is a reasonable approximation, and we can always include quadratic or cubic trends in the analysis if that is needed. Figure 10.1 shows the effect of the correction for attenuation for $d=0.5$ (medium effect size) and a reliability ranging between zero and one. It is clear that for reliabilities larger than 0.5 the relationship is almost linear.

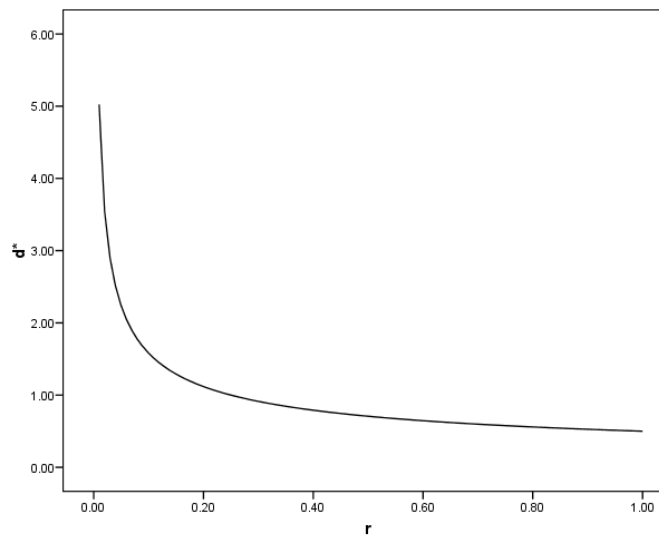


Figure 10.1 Corrected values for $d=0.5$ for a range of reliabilities.

The advantage of adding reliability as a predictor variable is that the effect of unreliability on the study outcomes is estimated based on the available data and not by a priori corrections. Another advantage is that we can test statistically if the correction has indeed a significant effect. Lastly, an interesting characteristic of multilevel modeling in meta-analysis is that it is possible to add an explanatory variable only to the random part, excluding it from the fixed part. Hence, if we suspect that a certain covariate, for instance poor experimental design, has an effect on the variability of the outcomes, we have the option to include it only in the random part of the model, where it affects the between studies variance, but not the average outcome.

A variation on correcting for artifacts is controlling for the effect of study size. An important problem in meta-analysis is the so-called *file drawer problem*. The data for a meta-analysis are the results from previously published studies. Studies that find significant results may have a greater probability to get published. As a result, a sample of published studies can be biased in the direction of reporting large effects. In classical meta-analysis, one way to investigate this issue is to carry out a fail-safe analysis (Rosenthal, 1991). This answers the question how many unpublished insignificant papers must lie in various researchers' file drawers to render the combined results of the available studies insignificant. If the fail-safe number is high, we assume it is unlikely that the file drawer problem affects our analysis. An alternative approach to the file drawer problem is drawing a *funnel plot*. The funnel plot is a plot of the effect size versus the total sample size (Light & Pillemer, 1984). Macaskill, Walter and Irwig (2001) recommend using the inverse of the sampling variance instead of the studies' sample size, because this is a more direct indicator of a study's expected variability; Sterne, Becker & Egger (2005) suggest using the standard error. These are all indicators of the studies' precision, and are highly correlated. If the sample of available studies is 'well-behaved' the plot should be symmetric and have the shape of a funnel. The outcomes from smaller studies are more variable, but estimate the same underlying population parameter. If large effects are found predominantly in smaller studies, this indicates the possibility of publication bias, and the possibility of many other non-significant small studies remaining unpublished in file drawers. In addition to a funnel plot, the effect of study sample size can be investigated directly by including the total sample size of the studies as an explanatory variable in a multilevel meta-analysis. This variable should *not* be related to the outcomes. When instead of sample size the standard error of the effect is included in the model as a predictor, the resulting test is equivalent to the Egger test, a well-known test for funnel asymmetry (Sterne & Egger, 2005).

The example data in Table 10.2 have an entry for the reliability of the outcome measure (r_{ii}). These (fictitious) data on the effect of social skills training assume that two different instruments were used to measure the outcome of interest; some studies used one instrument, some studies used another instrument. These instruments, in this example tests for social anxiety in children, differ in their reliability as reported in the test manual. If we use classical psychometric methods to correct for attenuation by unreliability, followed by classical meta-analysis using the random-effects model, the combined effect size is estimated as 0.64 instead of the value of 0.58 found earlier. The parameter variance is estimated as 0.23 instead of the earlier value of 0.17.

Table 10.4 Multilevel meta-analyses on example data

Model:	intercept- only	+	+	+	+
		N_{tot}	reliability	duration	all
intercept	0.58 (.11)	0.58 (.11)	.58 (.11)	0.57 (.08)	0.58 (.08)
N_{tot}		0.001 (.01)			-.00 (.01)
reliability			.51 (1.40)		-.55 (1.20)
duration				0.14 (.04)	0.15 (.04)
σ_u^2	0.14	0.16	0.16	0.04	0.05
p -value χ^2 deviance test	$p < .001$	$p < .001$	$p < .001$	$p = .15$	$p = .27$
p -value χ^2 residuals test	$p < .001$	$p < .001$	$p < .001$	$p = .09$	$p = .06$

If we include the reliability and the sample size as explanatory variables in the regression model, we obtain the results presented in Table 10.4. The predictor variables are centered on their grand mean, to retain the interpretation of the intercept as the ‘average outcome’. The first model in Table 10.4 is the empty ‘intercept only’ model presented earlier. The second model, which follows equation (10.2), includes the total sample size as a predictor. The third model includes the reliability of the outcome measure. The fourth model includes the duration of the experiment, and the fifth includes all available predictors. Both the univariate and the multivariate analyses show that only the duration has a significant effect on the study outcomes. Differences in measurement reliability and study size are no major threat to our substantive conclusion about the effect of duration. Since there is no relation between the study size and the reported outcome, the existence of a file drawer problem is unlikely.

The last model that includes all predictor variables simultaneously is instructive. The (non-significant) regression coefficient for reliability is negative. This is counterintuitive. This is also in the opposite direction of the regression coefficient in the model (3) with reliability as the only predictor. It is the result of a so-called ‘repressor’ effect caused by the correlations (from 0.25 to 0.33) among the predictor variables. Since in meta-analysis the number of available studies is often small, such effects are likely to occur if we include too many explanatory study-level variables. In the univariate model (10.3), the regression coefficient of reliability is 0.51. This implies that, if the reliability goes from 0.75 to 0.90, the expected outcome increases by $(0.15 \times 0.51 =) 0.08$. This is reasonably close to the correction of 0.06 that results from applying the classical correction for attenuation. However, the large standard error for reliability in model (10.3) suggests that this correction is not needed. Thus, the corrected results using classical methods may well be misleading.

In meta-analysis it is typical to have many study characteristics— and typically many of these are correlated. This leads to substantial multicollinearity, and makes it difficult to determine what effects are important. The approach taken above, to evaluate each effect separately and next look at multiple effects, is a reasonable strategy. The problem of predictor variable selection is a general problem in multiple regression when there are many potential predictors, but it is especially important in meta-analysis because the number of available studies is often small.

10.5 MULTIVARIATE META-ANALYSIS

The example in section 10.4 assumes that for each study we have only one effect size, which leads to analysis models with two levels. However, there are several situations that can lead to three-level models. Three-level structures are appropriate if there are multiple studies within the same publication (or multiple studies by the same group of researchers), or if there are multiple effect sizes used in the same study. Such situations lead to a meta-analysis with multiple effect measures, sometimes denoted as a multiple endpoint meta-analysis (Gleser & Olkin, 1994). Three-level structures are also appropriate if the studies investigate the difference between several different treatment groups and one control group. This leads to a collection of effect size measures, which all share the same control group, sometimes denoted as a multiple treatment meta-analysis (Gleser & Olkin, 1994). In both cases, there are dependencies between the effect sizes within studies.

In classical meta-analysis, such dependencies are often ignored by carrying out a series of univariate meta-analyses, or solved by calculating an average effect size across all available outcome measures. For several reasons, this approach is not optimal, and more complex procedures have been proposed to deal with dependent effect sizes (Gleser & Olkin, 1994).

In a multilevel model, we can deal with multiple dependent effect sizes by specifying a multivariate outcome model. Thus, a level is added for the multiple outcome variables, analogous to the multivariate multilevel models discussed in Chapter Nine. When some studies do not report on all available outcomes, we have a missing data problem, which is dealt with in the same way as in a standard multivariate multilevel model.

The univariate model for meta-analysis is written as $d_j = \gamma_0 + u_j + e_j$ (cf. equation 10.4). The corresponding equation for a bivariate random effects meta-analysis is

$$d_{ij} = \gamma_{0j} + u_{ij} + e_{ij} \quad (10.11)$$

In equation 10.11, the j is an index that refers to the outcome, and in the bivariate case $j = 1, 2$. The sampling variances of e_{i1} and e_{i2} are assumed known, and in addition the covariance between the sampling errors e_{i1} and e_{i2} is assumed known. The variances and covariance of the u_{ij} that represent differences between studies are estimated. Thus, replacing the variance terms in the univariate meta-analysis, we have the known covariance matrix Ω_e at the second level, and the estimated covariance matrix Ω_u at the third level. The lowest level is used only to specify the multivariate structure, following the procedures explained in Chapter Nine. Thus, in the bivariate case we have

$$\Omega_e = \begin{pmatrix} \sigma_{e11}^2 & \sigma_{e1e2} \\ \sigma_{e1e2} & \sigma_{e22}^2 \end{pmatrix} \tag{10.12}$$

and

$$\Omega_u = \begin{pmatrix} \sigma_{u11}^2 & \sigma_{u1u2} \\ \sigma_{u1u2} & \sigma_{u22}^2 \end{pmatrix} . \tag{10.13}$$

The covariance between e_1 and e_2 can also be written as $\sigma_{e1e2} = \sigma_{e1}\sigma_{e2}\rho_w$, where ρ_w is the known within study correlation. Generally, ρ_w is estimated by the correlation between the outcome variables in the control group or the pooled correlation across the control and experimental group (Gleser & Olkin, 1994). From the estimated matrix Ω_u we can calculate the between studies correlation ρ_B , using $\rho_B = \sigma_{u1u2} / \sigma_{u1}\sigma_{u2}$.

Table 10.5 Some effect measures, their transformation and sampling covariance

Measure	Estimator	Transformation	Sampling covariance
mean	\bar{x}	-	s^2/n
diff. 2 means	$g = (\bar{y}_E - \bar{y}_C) / s$	$d = (1-3/(4N-9))g$	$\rho_w(n_E+n_C)/(n_E n_C) + \rho_w^2 d_1 d_2 / (2(n_E+n_C))$
stand. dev,	S	$s^* = LN(s) + 1/(2df)$	$\rho_w^2 / (2df)$
correlation	R		$\sigma(r_{sb}, r_{sw}) = [0.5 r_{sb} r_{sw} (r_{su}^2 + r_{sv}^2 + r_{tu}^2 + r_{tv}^2) + r_{su} r_{tv} + r_{sv} r_{tu} - (r_{st} r_{su} r_{sv} + r_{ts} r_{tu} r_{tv} + R_{us} r_{uv} r_{ur} + r_{vs} r_{vru} r)] / n$
proportion	P	logit = $LN(p/1-p)$	$1 / ((np(1-p_1)(np(1-p_2)))$

Table 10.5 presents the formula for the sampling covariance of some effect measures (cf. Raudenbush & Bryk, 2002). The covariance between two correlations is a complicated expression discussed in detail by Steiger (1980) and presented in an accessible manner by Becker (2007).

Currently HLM is the only software that directly inputs the vector of effect sizes and the sampling (co)variances, although other software can be used with special command setups. These and other software issues are discussed in the appendix to this chapter.

A serious limitation for multivariate meta-analysis is that the required information on the correlations between the outcome variables is often not available in the publications. Some approximations may be available. For instance, if standardized tests are used, the test manual generally provides information on the correlations between subtests. If a subset of the studies reports the relevant correlations, they can be meta-analyzed in a preliminary step, to obtain a global estimate of the correlation between the outcomes. Riley, Thompson and Abrams (2008) suggest to set the within study covariance equal to the covariance between the effect measures. Alternatively, researchers can conduct a ‘sensitivity analysis’ in which a range of plausible values for the correlation between outcome measures can be used to determine the likely effect of this issue on substantive interpretations. Cohen (1988) suggested to use the value of 0.10 for a small correlation, 0.30 for a medium correlation, and 0.50 for a large correlation. Taking these suggestions as a starting point, a sensitivity analysis using the values 0.00, 0.10, 0.30 and 0.50 appears reasonable.

Table 10.6 Selected data from studies reporting odds ratios for asthma and LRD

ID	Size	Age	Year	Smoke	LOR	SE LOR	LOR	SE LOR
					asthma	asthma	Lrd	Lrd
3	1285	1.1	1987	0			0.39	0.27
4	470	9.0	1994	0	0.04	0.20		
6	1077	6.7	1995	0			0.35	0.15
8	550	1.7	1995	0	0.61	0.18		
10	850	9.4	1996	0			0.25	0.23
11	892	10.9	1996	0			-0.02	0.22

Table 10.6 Selected data from studies reporting odds ratios for asthma and LRD

ID	Size	Age	Year	Smoke	LOR	SE LOR	LOR	SE LOR
					asthma	asthma	Lrd	Lrd
14	1232	9.5	1996	0			-0.09	0.27
16	3048	1.0	1997	0	0.42	0.12		
17	2216	8.6	1997	1			-0.27	0.15
19	535	6.5	1995	0			0.30	0.60
20	5953	0.5	1987	0	0.99	0.20		
22	159	1.0	1986	0	0.69	0.24		
24	9670	5.0	1989	0	0.05	0.09	-0.04	0.12
25	318	8.2	1995	0			0.34	0.36
26	343	9.5	1995	0	0.85	0.28		
28	11534	9.5	1996	1	0.12	0.06	-0.02	0.11
29	10,106	7.5	1984	0	-0.02	0.06		
32	443	3.8	1992	0	1.93	0.46		
36	12727	2.5	1987	0	0.38	0.10		
37	257	7.5	1989	1	0.99	0.37		
38	511	2.5	1995	0			0.52	0.17
40	253	1.0	1995	0	-0.33	0.51		
43	1503	13.0	1997	0	0.12	0.14		
44	7677	14.0	1995	0	0.14	0.02	0.16	0.06
49	1001	7.0	1988	1	0.04	0.19		
50	961	13.5	1995	0	0.02	0.13		
51	1138	9.0	1983	0	0.01	0.11		
52	5412	9.2	1997	1	0.22	0.09	0.30	0.09
54	925	7.0	1997	1	0.53	0.32	0.47	0.28
56	2554	9.5	1995	1			-0.56	0.33
57	4665	13.5	1997	1	0.04	0.07		
59	1501	9.5	1991	1	0.18	0.15	0.10	0.11
61	1143	1.0	1981	0	0.04	0.01		
63	483	1.0	1989	0	0.48	0.23		
65	8154	4.6	1995	0	0.20	0.04		
69	2077	0.5	1974	0	0.50	0.21		
71	8585	7.0	1993	0			0.19	0.04
75	301	1.0	1988	1	0.69	0.35		
76	153	10.9	1989	0	0.92	0.39		
78	3072	8.5	1982	0			0.40	0.16
79	4331	2.5	1990	0			0.74	0.23
80	1215	2.0	1993	1			0.79	0.21
81	708	10.5	1994	1			0.34	0.19
82	620	10.0	1995	0			1.03	0.39
83	126	9.0	1991	0			0.94	0.37
84	228	8.5	1992	1			0.69	0.29
85	914	3.5	1993	0			0.36	0.18
88	16562	8.0	1991	0	0.30	0.30		
89	15712	5.0	1995	0	0.29	0.19		
93	3482	8.4	1986	0	0.22	0.08	0.14	0.13
105	192	0.8	1992	0	1.19	0.50		
109	726	16.3	1996	0	1.47	0.47		
111	372	7.0	1980	0	1.77	1.03		
113	684	9.0	1988	1	0.22	0.11	-0.07	0.04
114	207	9.3	1992	1			0.36	0.33
122	700	7.9	1992	0			0.35	0.19
601	9653	9.0	1997	1			0.44	0.14
603	3769	9.0	1997	1			0.10	0.23
603	2540	9.0	1997	1			-0.36	0.43

To illustrate the proceedings we analyze a bivariate meta-analytic data set discussed by Nam, Mengersen and Garthwaite (2003). The data are from a set of studies that investigate the relationship between children's environmental

exposure to smoking (ETS) and the child health outcomes asthma and lower respiratory disease (LRD). Table 10.6 lists the logged odds-ratio (LOR) for asthma and LRD, and their standard errors. Study level variables are average age of subjects, publication year, smoking (0 parents, 1 other in household), and covariate adjustment used (0=not, 1=yes).

There are two effect sizes, the logged odds ratio for asthma and lower respiratory disease (LRD). Only a few studies report both. There is no information on the correlation between LOR asthma and LRD within the studies, at the study-level the correlation is 0.80. However, this is an ecological correlation, which confounds within study and between study effects. The analysis choices are to model the within study variances only, setting the covariance to zero, or to assume a common value for the within study variance (for example, $r = 0.30$, a medium size correlation). As a first approximation, we set the covariance to zero. To analyze these data, the data file in Table 10.6 must be restructured in a 'long' or 'stacked' file format (cf. Chapter Nine), where the outcomes for asthma and LRD become conditions i nested within studies j . Since many studies report only one outcome, these studies have only one condition. The intercept-only model for these bivariate data is

$$LOR_{ij} = \beta_{0j} Astma + \beta_{1j} LDR + e(A)_{ij} + e(L)_{ij}. \quad (10.14)$$

In equation 10.14, the variables *Astma* and *LRD* are dummy variables referring to the asthma and LDR outcomes. The error terms $e(A)_{ij}$ and $e(L)_{ij}$ are also specific for each outcome, and are assumed to be uncorrelated.

Table 10.7 Results bivariate meta-analysis for exposure to smoking, covariance between outcomes constrained to zero

Model	Interc. only	Equality	+ Age (centered)
Fixed part	Coeff. (s.e.)	Coeff. (s.e.)	Coeff. (s.e.)
Interc. Asthma	0.32 (.04)	0.29 (.04)	0.29 (.03)
Interc. LRD	0.27 (.05)	0.29 (.04)	0.29 (.03)
Age			-0.03 (.006)
Random part			
Var (Asthma)	0.06 (.02)	0.08 (.03)	0.06 (.02)
Var (LRD)	0.07 (.02)	0.05 (.02)	0.03 (.01)
Covar (AL)	0.06 (.02)	0.06 (.02)	0.05 (.01)
Deviance	44.2	44.7	32.4

Table 10.7 presents the results of a series of models for the exposure to smoking data. The intercept-only model shows a clear effect; children who are exposed to environmental smoking face increased odds of having asthma or LRD. The variances are significant by the Wald test. Both the univariate deviance difference test (removing the study level variances individually) and the multivariate deviance difference test (removing the study level variances simultaneously) confirm this. The effect on LRD appears somewhat stronger than the effect on asthma. In a multivariate analysis this difference can be tested by a Wald test, or by constraining these regression coefficients to be equal. The Wald test for this equality constraint produces a chi-square of 1.232, with $df=1$ and $p=0.27$ clearly not significant. The column marked 'equality' in Table 10.7 reports the estimates when an equality constraint is imposed on the regression coefficients for asthma and LRD. The deviance difference test can not be used here to test the difference, since the estimates use REML. The last column adds the variable Age, which is the only significant study level predictor. Older children show a smaller effect of exposure to smoking. Age is entered as a single predictor, not as interactions with the asthma or LRD dummy. This assumes that the effect of age on Asthma and LRD is the same. Exploratory analysis show indeed very similar regression coefficients when asthma and LRD are modeled separately in a bivariate meta-analysis, and the Wald equality test on the regression coefficients is not significant ($p=.45$).

The results reported in Table 10.7 are estimates where the common correlation between the outcomes is constrained to be zero. This causes some bias; Riley, Thompson and Abrams (2008) report a simulation that shows that this leads to an upward bias in the study-level variances, which in turn leads to some bias in the fixed effects and increased standard errors for the fixed effects. They recommend to either impute a reasonable value for r , or to estimate only one covariance parameter that confounds the within and between study level covariance. They report simulations that show that the latter strategy is quite successful. In our case, we carry out a sensitivity analysis, where several plausible values are specified for the covariance between the error terms $e(A)_{ij}$ and $e(L)_{ij}$. Since $e(A)_{ij}$ and $e(L)_{ij}$ are standardized to have a variance equal to one, the covariance is equal to the correlation. Table 10.8 shows the results when the common correlation is constrained to 0.1, 0.3 and 0.5 (Cohen's suggestions for a small, medium and large correlation) and when a single common covariance is estimated for the between study and the within study part. Two conclusions are evident. First, the estimated effect of passive smoking on asthma and LRD is similar to the results reported in Table 10.7, and second, all results are remarkably similar. It should be noted that in this meta-analytic data the variation between studies is small. Riley et al. (2008) report simulations that show that with larger between study variation the differences are larger. Still, they also report that the estimates for the fixed effects are not affected much by the different specifications for the within study correlation.

Table 10.8 Results bivariate meta-analysis for exposure to smoking, for 3 values of covariance between outcomes

Covariance =	0.10	0.30	0.50	common
Fixed part	Coeff. (s.e.)	Coeff. (s.e.)	Coeff. (s.e.)	Coeff. (s.e.)
Interc. Asthma	0.29 (.03)	0.29 (.03)	0.29 (.03)	0.29 (.03)
Interc. LRD	0.29 (.03)	0.29 (.03)	0.29 (.03)	0.29 (.03)
Age	-.03 (.006)	-.03 (.006)	-.03 (.006)	-.03 (.006)
Random part				
Var (Asthma)	0.03 (.01)	0.03 (.01)	0.03 (.01)	0.03 (.01)
Var (LRD)	0.06 (.02)	0.06 (.02)	0.07 (.02)	0.06 (.02)
Covar (AL)	0.05 (.01)	0.05 (.01)	0.05 (.01)	0.05 (.01)
Deviance	32.6	32.3	31.6	32.7

Multivariate meta-analysis is especially useful if most studies do not report on all possible outcomes. A series of univariate meta-analyses on such data assumes that the missing outcome variables are missing completely at random. A multivariate meta-analysis assumes that the missing outcomes are missing at random, a less strict assumption. In addition, a multivariate meta-analysis allows testing equality of effect sizes and regression coefficients, as exemplified in the bivariate exposure to smoking example.

For details on multivariate multilevel meta-analysis see Kalaian and Raudenbush (1996), Normand (1999), van Houwelingen, Arends and Stijnen (2002) and Kalaian and Kasim (2008). An example of a multivariate multilevel meta-analysis is discussed by Berkey et al. (Berkey, Hoaglin, Antczak-Bouckoms, Mosteller & Colditz, 1998).

Another interesting extension of multilevel meta-analysis arises when we have access to the raw data for at least some of the studies. This situation leads to a multilevel model that combines both sufficient statistics, as in standard meta-analysis, and raw data to estimate a single effect size parameter. Higgins et al. (Higgins, Whitehead, Turner, Omar & Thompson, 2001) describe the general framework for this hybrid meta-analysis, and discuss classical and Bayesian analysis methods. Examples of such hybrid meta-analyses are the studies by Goldstein et al. (Goldstein, Yang, Omar, Turner & Thompson, 2000) and Turner et al. (Turner, Omar, Yang, Goldstein & Thompson, 2000).

10.6 STATISTICAL AND SOFTWARE ISSUES

The program HLM (Raudenbush et al., 2000) has a built-in provision for meta-analysis, which is restricted to two-levels. If we need three levels, we can use the standard HLM/3L software, using an adapted program setup. Other software can be used, provided it is possible to put restrictions on the random part. MLwiN (Rasbash et al., 2000) and Proc Mixed in SAS (Littell et al., 1996) all have this capacity, and can therefore be used for meta-analysis, again with an adapted setup. Ways of tweaking HLM and MLwiN for meta-analysis are discussed in the Appendix to this chapter.

There are some minor differences between the programs. HLM uses by default an estimator based on Restricted Maximum Likelihood (RML), while MLwiN by default uses Full Maximum Likelihood (FML, called IGLS in MLwiN). Since RML is theoretically better, especially in situations where we have small samples and are interested in the variances, for meta-analysis we should prefer RML (called RIGLS in MLwiN). If in a specific case the difference between RML and FML is small, we can choose FML because it allows testing regression coefficients using the deviance difference test. The results reported in this chapter were all estimated using RML.

An important difference between HLM and other multilevel analysis software is the test used to assess the significance of the variances. HLM by default uses the variance test based on a chi-square test of the residuals (Raudenbush & Bryk, 2002, cf. Chapter Three of this book). MLwiN estimates a standard error for each variance, which can be used for a Z-test of the variance. In meta-analysis applications, this Z-test is problematic. Firstly, it is based on the assumption of normality, and variances have a chi-square distribution. Secondly, it is a large-sample test, and with small sample sizes and small variances the Z-test is inaccurate. In meta-analysis the sample size is the number of studies that are located, and it is quite common to have at most 20 studies. An additional advantage of the chi-square test on the residuals is that for the empty model this test is equivalent to the chi-square variance test in classical meta-analysis (Hedges & Olkin, 1985). The variance tests reported in this chapter use both the deviance difference test and the chi-square test on the residuals. MLwiN does not offer this test, but it can be produced using the MLwiN macro language.

It should be noted that the standard errors that are used to test the significance of the regression coefficients and to establish confidence intervals are also asymptotic. With the small samples common in meta-analysis, they can lead to confidence intervals that are too small, and p -values that are spuriously low (Brockwell & Gordon, 2001). It appears prudent not to use the standard normal distribution, but the Student t -distribution with degrees of freedom equal to $k-p-1$, where k is the number of studies and p the number of study-level explanatory variables in the model. In HLM this is the standard test for the regression coefficients. In simulations by Berkey et al. (Berkey et al., 1998) this provided correct p -values. Brockwell and Gordon (2001) recommend profile likelihood methods and bootstrapping. These are treated in Chapter XXX in this book.

For estimating complex models, Bayesian procedures are promising and are coming into use (cf. Sutton et al., 2000). These use computer-intensive methods such as Markov Chain Monte Carlo (MCMC) to estimate the parameters and their sampling distributions. These methods are attractive for meta-analysis (DuMouchel 1980, 1994; Smith,

Spiegelhalter & Thomas, 1995) because they are less sensitive to the problems that arise when we model small variances in small samples. Bayesian models are treated in Chapter XXX in this book. Bayesian modeling starts with the specification of a prior distribution that reflects a-priori beliefs about the distribution of the parameters. In principle, this is an elegant method to investigate the effect of publication bias. An example of such an analysis is Biggerstaff, Tweedy and Mengersen (1994). Although the software MLwiN includes Bayesian methods, at present these cannot analyze meta-analytic models, and more complicated software is needed, such as the general Bayesian modeling program BUGS (Spiegelhalter, 1994). Nam, Mengersen and Garthwaite (2003) discuss several Bayesian meta-analysis models, using the exposure to smoking data as their example.

APPENDIX

Software Implementation: General

Multilevel meta-analysis requires software that allows imposing constraints on the random part. If that is possible, there are two different but functionally equivalent methods to specify the known variance. The first method is to add the standard error of the outcome to the model as a predictor. This predictor should be in the random part at the first level only, and not in the fixed part or elsewhere in the random part. Next, the variance for this predictor is constrained equal to one. The second method is to use the inverse of the sampling variance (the square of the standard error) as a weight at the first level. The weights must be raw weights, not normalized. Next, the variance at the lowest level is constrained to one. The weight method is the most general, for instance, Cheung (2008) uses it to carry out a meta-analysis in the structural equation software Mplus.

In HLM and MLwiN both methods are available. To apply multilevel models in meta-analysis in other software, such as SAS Proc Mixed, the software must have options to set up a model using constraints as specified for MLwiN or for HLM/3L. This means that it must be possible to have a complex lower-level variance structure, as in MLwiN, or to constrain the lowest-level variance to 1 and to add a weight variable, as in HLM/3L. So far, public domain software for multilevel analysis does not offer these options. For univariate meta-analysis, David Wilson (Lipsey & Wilson, 2001) has written macros for SAS, SPSS and STATA that carry out both ML and RML based meta-regression, equivalent to the multilevel approach.

Software Implementation: HLM

The simplest program for multilevel meta-analysis is the meta-analysis module that is part of the software HLM. HLM expects for each study a row of data containing the study ID, an effect measure, and its sampling variance, followed by the explanatory variables. This software can also carry out a multivariate meta-analysis, but does use listwise deletion on the studies that have missing outcome variables.

HLM can also be used for meta-analysis using the weight method described earlier. In HLM, the weight that is supplied is the sampling variance itself, and the estimation specification must include a specification that the weights are sampling variance weights. From that point, the analysis proceeds as a standard HLM analysis.

If we need more than two levels in HLM, we must use HLM/3L, using the sampling variance weight method. Multivariate multilevel meta-analysis is done using a separate level for the outcome variables, as illustrated earlier.

HLM can be tweaked to analyze a multivariate meta-analysis using the sampling variance weighting method. This specifies the within study variance correctly, but assumes uncorrelated outcomes. Kalaian and Raudenbush (1996) describe a method to transform the multivariate outcome matrices to make the outcomes orthogonal, without disturbing the parameter estimates for the fixed part.

Software Implementation: MLwiN

Using MLwiN is more complicated. The data structure is analogous to HLM: we need a study ID, the effect size, its standard error (the square root of the sampling variance), the regression constant (HLM includes this automatically), and the explanatory variables. To set up the analysis, we distinguish two levels: the outcomes are the first level and the studies the second. Usually we have one outcome per study, so there is no real nesting. The predictor 'standard error' is included only in the random part on level 1, with a coefficient fixed at 1. The regression constant is included in the fixed part and in the random part at level 2. Explanatory variables are included in the fixed part only. MLwiN does not produce the chi-square test on the variances. The formula for the chi-square test is $\chi^2 = \sum \left((d_j - \hat{d}_j) / s.e.(d_j) \right)^2$, which is

the sum of the squared residuals divided by their standard errors. The degrees of freedom are given by $df = J - q - 1$, where J is the number of studies, and q the number of explanatory variables in the model. Assuming that the outcomes are denoted by ' d ', and the standard errors of the d 's by ' sed ', the sequence of MLwiN commands for computing the chi-square is:

- PRED C50 (produce predicted values)
- CALC C50=((d -C50)/ sed)^2 (produce squared residuals/se)
- SUM C50 to B1 (sum to box B1)

- CPRO B1 *df* (calculate *p*-value).

This code assumes that the spreadsheet column C50 is unused.

In a multivariate meta-analysis the setup is more complicated. There are multiple outcomes, with each outcome indicated by a 0-1 dummy variable that indicates that outcome. There is no intercept in the model, so for *p* outcomes all *p* dummy variables can be used. This structure can be generated automatically in MLwiN using the multiple response option. There are as many standard error variables as there are outcome variables. To accomplish this, the standard error is multiplied by each of the dummy variables, which produces a column with standard errors for that specific outcome, and zeros elsewhere. MLwiN can do this automatically when the multiple response option is activated. The standard error variables are all entered only in the random part at the lowest level, with their variances constrained to one. The covariances are constrained zero or to their common known or imputed value. In a multivariate meta-analysis the residuals chi-square must be calculated for each outcome separately. For example, for the bivariate meta-analysis on the log odds ratio (LOR) in the exposure to smoking example, the sequence of MLwiN commands is:

- PRED C50
- CALC C50=((‘LOR’-C50)/ ‘SE_LOR’)^2
- CALC C51=C50*‘ASTMADUMMY’
- SUM C51 TO B1
- CPRO B1 *df*
- CALC C52=C50*‘LRDDUMMY’
- SUM C52 TO B2
- CPRO B2 *df*

Multiplication of the values in C50 by the asthma dummy and the LRD dummy creates a column C51 and C52 that contains only values for asthma and LRD respectively. These are summed separately. In the exposure to smoking example, this produces a value for chi-square of 148.6 for asthma and 1041.0 for LRD. For asthma we have $df=32-1=31$ and for LRD we have $df=35-1=34$. Both chi-square values are highly significant.

MLwiN can also use the inverse sampling variance weight method. In a multivariate meta-analysis, the standard errors as variables method is preferred, because then we can specify a correlation between the outcomes. In the inverse sampling weight method, this correlation is assumed to be zero.