

5

Multilevel Models for Meta-Analysis

Joop J. Hox Edith D. de Leeuw
Utrecht University, Utrecht Methodika, Amsterdam

Meta-analysis concerns the statistical integration of a large number of results from empirical studies (cf. Glass, 1976). The goal is to summarize the results of a collection of independently conducted studies on one specific research question. 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 on integrating or summarizing the outcomes, but on more detailed questions about variations in the outcomes, such as: What is the effect of different durations for the training sessions? Are there differences between different training methods? In these cases, the meta-analyst not only examines the overall study outcomes, but also codes study characteristics. These study characteristics, for example design features or type of subjects sampled, are potential explanatory variables to explain differences in the study outcomes.

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. A very general model for meta-analysis is the random effects model (Hedges & Olkin, 1985). In this model, the focus is on analyzing the size of the effects found in the different studies, not on establishing the statistical significance of a combined outcome. The random effects model assumes that study outcomes vary not only because of

random sampling effects (variations within each study), but also because of real differences between the studies. For instance, study outcomes can vary because different studies employ different sampling methods, use different experimental manipulations, or measure the effects with different instruments. The random effects model is used to decompose the variance of the study outcomes into a part that is the result of sampling variation, and a part that reflects real differences between the studies. Hedges and Olkin (1985) gave procedures that can be used to decompose the total variance of the study outcomes into random sampling variance and systematic between-study variance, and to test the significance of the between-study variance. If the between-study variance is large and significant, the study outcomes are *heterogeneous*. In that case, the usual procedure is to form clusters of studies that differ in their outcomes, but that are homogeneous within the clusters. These clusters can be constructed a priori, based on available study characteristics; they can also be constructed a posteriori, based on a cluster analysis of the reported outcomes. The goal is to identify study characteristics that explain differences between the study outcomes. Variables that affect the study outcomes are in fact moderator variables, that is, variables that interact with the independent variable.

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 was 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, 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 grouping variable. The general multilevel regression model assumes that each study has its own regression model. In particular, the intervention effect (e.g., the regression coefficient for the grouping variable in each study) is allowed to vary across studies. Standard multilevel analysis can be used to estimate the mean and variance of the intervention effects across the studies. If the intervention effects vary substantially and significantly across studies, we have heterogeneous results. In that case, we can study further the variation of intervention effects across studies by examining study-level regression models that attempt to explain the study-specific intervention effects with the available study characteristics as explanatory variables.

These analyses can be carried out using standard multilevel regression methods (cf. Bryk & Raudenbush, 1992; Goldstein, 1995; Hox, 1995; Snijders & Bosker, 1999) and standard multilevel software. A special

complication is that 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. Hunter and Schmidt (1990) discussed these methods in detail, and Hedges and Olkin (1985) discussed the statistical models used.

Nevertheless, it is possible to carry out a multilevel meta-analysis on the data that are usually available in meta-analysis. Raudenbush and Bryk (Bryk & Raudenbush, 1992; Raudenbush & Bryk, 1985) presented 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. Analytic procedures for the standard multilevel software HLM and MlwiN are given in the Appendix. The major advantage of using multilevel analysis instead of classical meta-analysis methods is flexibility. 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 MODEL

In a typical meta-analysis, the studies usually employ 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, such as a correlation coefficient or the standardized difference between two means, d . The general model for the study outcomes is

$$d_j = \delta_j + e_j. \quad (5.1)$$

In Equation 5.1, d_j is the outcome of study j ($j=1, \dots, J$), δ_j is the population value of this outcome, and e_j is the sampling error for this 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 20 (Hedges & Olkin, 1985) to 30 (Bryk & Raudenbush, 1992), the assumption that the sampling distribution of the outcomes is normal is usually reasonable. Most classical meta-analysis methods also assume normality (cf., Hedges & Olkin, 1985). The variance of the sampling distribution is often known from statistical theory; in some cases, a transformation is needed to achieve normality and known sampling

TABLE 5.1
Effect Measures and Their Sampling Variance

Measure	Estimator	Transformation	Sampling variance
Mean	\bar{X}	-	$\frac{s^2}{n}$
Diff. 2 Means	$g = \frac{(\bar{Y}_E - \bar{Y}_C)}{s}$	-	$\left(\frac{n_E + n_C}{n_E n_C}\right) + \left(\frac{g^2}{2n_E + n_C}\right)$
St. Dev.	s	$s^* = LN(s) + \frac{1}{2}df$	$\frac{1}{2}df$
Correlation r		$z = .5LN\left[\frac{(1+r)}{(1-r)}\right]$	$1/(n-3)$
Proportion p		$1^* = LN[p/(1-p)]$	$1/[np(1-p)]$

variance. Table 5.1 lists some common effect size measures; if needed, the normalizing transformation; and the corresponding sampling variance.

When using the sample mean as the effect measure, we should make sure that the outcomes are comparable across studies. If different outcome measures are used, the measures might be scaled in different units. Without some kind of standardization, comparing those outcomes is like comparing apples and oranges, or rather, like comparing pounds and kilograms.

In Table 5.1, g is the effect size proposed by Glass (1976). The transformation of s to s^* for the standard deviation is proposed by Bryk and Raudenbush (1992). The transformation for the correlation r is the familiar Fisher- Z transformation, and for the proportion, the logit. Usually, if a confidence interval is constructed for the transformed variable, the endpoints are translated back to the original estimator. For a more extended list of effect size measures and their sampling variance, see Rosenthal (1994) and Cornell and Mulrow (1999).

Equation 5.1 shows that the parameters δ_j , the study outcomes, are assumed to vary across the studies. The variation of δ_j is assumed to follow the regression model

$$\delta_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \dots + \gamma_p Z_{pj} + u_j, \quad (5.2)$$

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 . In meta-analysis, there are typically two kinds of study characteristics that can be used as explanatory variables in the regression model; methodological characteristics like study size, methodological quality, and reliability of instruments, and variables that are of theoretical interest, such as the type and intensity of intervention, or duration of the intervention.

By substituting Equation 5.2 into Equation 5.1 we get the complete model

$$d_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \dots + \gamma_p Z_{pj} + u_j + e_j. \quad (5.3)$$

If we have no explanatory variables, the model reduces to

$$d_j = \gamma_0 + u_j + e_j. \quad (5.4)$$

Equation 5.4, which in multilevel analysis is often denoted as the "intercept only" or "null" model (cf., Bryk & Raudenbush, 1992), is equivalent to the random effects model for meta-analysis described by Hedges and Olkin (1985). Hedges and Olkin described a one-step weighted least squares procedure for estimating the model parameters. Multilevel analysis programs typically use iterative maximum likelihood estimation, which, in general, is more efficient (Raudenbush, 1994). In practice, both models usually produce very similar parameter estimates.

In Equation 5.4, the intercept γ_0 is the estimate for the mean outcome \bar{d} across all studies. The variance of the outcomes (d_j) across studies, σ_u^2 , indicates how much the studies' outcomes vary. Thus, to test if the study outcomes are homogeneous is equivalent to testing the null hypothesis that σ_u^2 is equal to zero. If the test of σ_u^2 is significant, the study outcomes are heterogeneous.

Note that Equation 5.4 contains two residual error terms; u_j and e_j . The variance of the u_j , σ_u^2 , represents the true variation between the studies, which is estimated in the meta-analysis, and which we would like to explain using the available study characteristics. The e_j represents the differences between the studies that are the result of sampling variation. The sampling variance of the studies, $\sigma_{e_j}^2$, is determined fully by the within-study variation and sample size, and assumed known from the study's publications. Consequently, the sampling variance, calculated from the published results, is part of the data to be input in the program (software implementations for multilevel meta-analysis are discussed in an appendix to this chapter). Because the $\sigma_{e_j}^2$ are directly input as known data, there is no assumption that they are homogeneous, that is, that all $\sigma_{e_j}^2$ are equal. Typically, σ_e^2 , which is the (weighted) average sampling variance, is estimated by subtracting σ_u^2 from the total variance between studies. The proportion of between-study variance is estimated by the intraclass correlation, which can be estimated using the intercept-only (null) model as $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$. The proportion of systematic between-study variance can be used as an additional indicator of the degree of heterogeneity of the study outcomes. This is analogous to using the proportion of explained variance in standard regression models to

indicate the importance of specific predictor variables. Hunter and Schmidt (1990) pointed out that with a large number of studies, the power of the significance test is high, and small variances will become significant. When the number of studies is small, lack of significance for σ_u^2 does not imply that the outcomes are homogeneous. Hunter and Schmidt (1990) proposed a 25% rule of thumb: that is, if the between-study variance is 25% or more of the total variance, it is interesting enough to merit exploration. In our terminology, if the intraclass correlation ρ is 0.25 or higher, the variance between studies is deemed large enough to attempt to model it using the available study characteristics.

The general Equation 5.3 includes study characteristics Z_{pj} to explain differences in the studies' outcomes. In Equation 5.3, σ_u^2 is the residual between-study variance after the explanatory variables are included in the model. The 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 is still unexplained between-study variance left in the outcomes. The difference between the between-studies variance σ_u^2 in the null 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 included in Equation 5.3.

The multilevel meta-analysis model given by Equation 5.3 is similar to the general model for fixed effects as described by Hedges and Olkin (1985, chap. 8). In our notation, their model is given by

$$d_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + \dots + \gamma_p Z_{pj} + e_j. \quad (5.5)$$

Compared to Equation 5.3, Equation 5.5 lacks the study-level residual error term u_j . Thus, the general model for fixed effects described by Hedges and Olkin is a special case of the multilevel meta-analysis model. Omitting the study-level residual error term u_j implies the assumption that the explanatory variables in the model explain all of the variance across the studies. There are situations, for instance when we limit our statistical generalization to the set of studies at hand, where the fixed effects model is appropriate (Hedges & Vevea, 1998). If this is the case, we can model differences between studies using fixed effects (Overton, 1998). Alternatively, we may find empirically that the studies are homogeneous, meaning that the estimate for the between-study variance σ_u^2 is small and insignificant.

However, in general, we do want to generalize beyond the specific set of studies at hand, and experience shows that usually we cannot explain all between-study variance, just as in ordinary multiple regression analysis, we seldom find a multiple correlation coefficient equal to one. Hunter and Schmidt (1990) assumed that between-studies heterogeneity is partly

due to a large number of possible artifacts in the meta-analysis. An example of such an artifact is the (usually untestable) assumption of a normal distribution for the sampling errors e_j . However, unless the sample size is very small in some studies, the normality assumption for e_j is usually reasonable by central limit theorem. Other artifacts that may cause variation between the studies are the correctness of the statistical assumptions made in the original analyses, differences in the reliability of the instruments used in different studies, coder unreliability in coding study characteristics, and so forth. It is unlikely that the available study-level variables cover all of these artifacts. Generally, the amount of detail in the input for the meta-analysis, which are the research reports, papers, and articles, is not enough to cover all of these study characteristics. Therefore, to some extent, heterogeneous results are to be expected. It is this reasoning that led Hunter and Schmidt to their rule of thumb that the between-study variance in the null model should be larger than 25% of the total variance. The same reasoning led us to the conclusion that, in general, random effects models, such as multilevel regression models, should be used in meta-analysis.

Overton (1998) examined the differences between fixed and random effects models for meta-analysis using simulation. He found, not surprisingly, that fixed effects models perform best when data are generated following a fixed model, and random effects models perform best when data are generated following a random model. Because the fixed effects model is a special case of the random effects model, the best analysis strategy appears to be to begin by estimating a random effects model. If the between-study variance σ_u^2 turns out to be insignificant and negligible in size, the between-study variance can be fixed at zero, which effectively turns the multilevel analysis into a fixed effects analysis.

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 program META by Schwarzer (1989). This program is based on methods and procedures described by Rosenthal (1984), Hunter and Schmidt (1990), and Hedges and Olkin (1985). The (constructed) data set consists of 20 studies that compare an experimental group and a control group.¹

¹The example data were constructed using a regression model like Equation 5.4 with a single explanatory variable "duration in weeks", which was simulated from a normal distribution ($\mu = 6$, $\sigma = 2.5$). The population effect size δ for each study was predicted using a regression slope of 0.15 for the duration with mean outcome 0.6 across all studies.

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 preferred a corrected effect measure d : $d = \{1 - 3 / [4(n_E + n_C) - 9]\} g$. The sampling variance of the effect estimator d is equal to $(n_E + n_C) / (n_E n_C) + d^2 / [2(n_E + n_C)]$ (Hedges & Olkin, 1985, p. 86).

Table 5.2 lists both g and d for all 20 studies. With commonly used sample sizes, the difference between the two is very small. Table 5.2 also presents the sampling variance of d [$\text{var}(d)$], the one-sided p -value of the t test for the difference between 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 one study-level explanatory variable, the duration in number of weeks of the experimental intervention. It is plausible to assume that longer interventions lead to a larger effect. In Table 5.2, the studies are presented in increasing order of their effect sizes (g , d).

CLASSICAL META-ANALYSIS

Classical meta-analysis contains a variety of approaches that complement each other. An old approach is to combine the p values of the studies into one overall p value for the collection of studies. Several formulas are available for combining p -values. A popular procedure is the so-called Stouffer method (see Rosenthal, 1984). Each individual p is converted to the corresponding standard normal Z score. The Z scores are then combined using $Z = (\sum Z_i) / \sqrt{k}$, where Z_j is the Z score of study j , and k is the number of studies. For our example, the Stouffer method gives a combined Z of 7.73, which is highly significant ($p < 0.001$).

The combined p value gives us proof 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 used to combine the studies.

The sample sizes for the experimental and control group were generated independently from a normal distribution ($\mu = 30$, $\sigma = 10$), and the reliability was randomly chosen from the values 0.9 and 0.75. The reliability was used to attenuate the effect size δ , and finally, the observed effect size g was simulated by adding to δ , a random residual drawn from a normal distribution with mean 0 and variance determined by the sample sizes of the experimental and control groups.

TABLE 5.2
Example Results From Twenty Studies

Study	Duration	g	d	$var(d)$	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	.488	.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	29	.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

A meta-analysis of the effect sizes in Table 5.2, using the random effects model, estimates the overall effect as $\delta = 0.58$, with a standard error of 0.11. This gives us a Z value of 5.27 ($p < 0.001$). The 95% confidence interval for the overall effect size is $0.36 < \delta < 0.80$. The usual significance test of the variance is a chi-square test on the residuals, which for our example data leads to $\chi^2(19) = 48.9$, $p < 0.001$. As this 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 a (weighted) *average* of the distribution of effects in the population.

The between-study variance σ_u^2 is estimated as 0.17 and the proportion of between-study variance as 0.65. This is much larger than the 0.25 threshold that Hunter and Schmidt (1990) recommended for examining differences between studies. The usual approach in classical meta-analysis is to divide the studies into clusters that have different average effect sizes, while being internally homogeneous. A unidimensional cluster analysis of

the study outcomes can be used to form such groups. In our example, a cluster analysis produces three clusters. The first cluster consists of Studies 1 and 2, the second cluster consists of Studies 3 through 18, and the third cluster consists of Studies 19 and 20. For a post hoc interpretation of the clusters, we must examine how the clusters differ. If we look at the mean duration of the experiment in the three clusters, we find that this is 2 weeks in the first cluster, 6 weeks in the second, and 9 weeks in the third. Apparently, the duration of the experimental intervention indeed affects the study outcome.

Because we have the hypothesis that the duration of the experimental intervention is related to the outcome, we can also form a priori clusters based on this variable. We distinguish two a priori clusters; the first consists of the 9 studies that have a duration of 5 weeks or less, and the second consists of the 11 studies that have a duration of 6 weeks or more. The overall outcome in the first cluster is 0.33 ($SE = 0.15$, $p = 0.01$), and in the second cluster, 0.77 ($SE = 0.15$, $p < 0.001$). Studies with a longer duration have larger effect sizes. In both clusters, the null-hypothesis of homogeneous outcomes is rejected. The proportion of between-study variance is estimated as 0.61, $\chi^2(10) = 22.35$, $p = 0.01$ in the first cluster, and 0.69, $\chi^2(8) = 16.72$, $p = 0.03$ in the second cluster. We can perform a formal test for the difference between the two outcomes. Completely analogous to analysis of variance, where the total variance is partitioned into a between-groups variance and a within-groups variance, we can partition the total chi-square into a between-clusters chi-square and a within-clusters chi-square (Cooper, 1998; Hedges & Olkin, 1985). The total chi-square is the chi-square for the between-study variance for all 20 studies, which is $\chi^2(19) = 48.85$, $p < .001$. The within-clusters chi-square is given by the sum of the chi-squares for the variance within the two clusters, $\chi^2(18) = 39.07$, $p < 0.003$. The between-clusters chi-square is found by subtracting the within-clusters chi-square from the total chi-square; $\chi^2(1) = 9.78$, $p < = 0.002$. The between-clusters chi-square is highly significant ($p < 0.002$). The conclusion seems warranted that duration of experimental intervention has an effect on the outcome. However, the within-clusters chi-square was also significant. The fact that we still have significant heterogeneity in the two clusters indicates that we have not explained all systematic differences between the studies.

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 reported earlier. The intercept, which in the absence of other explanatory

TABLE 5.3
Results of Random Effects Model and Multilevel Regression Analyses on Example Data

<i>Model</i>	<i>Classical Random Effects</i>	<i>Multilevel Null Model</i>	<i>Multilevel Using Duration</i>
intercept	$\delta = 0.58$ (.11)	$\gamma_0 = 0.58$ (.11)	$\gamma_0 = 0.57$ (.08)
duration parameter	.17	.14	$\gamma_1 = 0.14$ (.03)
variance σ_u^2			.04
p value χ^2 test ^a	$p < .001$	$p < .001$	$p = .09$

^aChi-square test on residuals, cf. Hedges & Olkin, 1985, and Bryk & Raudenbush, 1992.

variables is the overall outcome that classical meta-analysis indicates by δ , is estimated as $\gamma_0=0.58$, with a standard error of 0.11 ($p < 0.001$). The null hypothesis of homogeneous outcomes is rejected, but the between-study variance is estimated a bit lower than in the classical meta-analysis. The between-study variance σ_u^2 is estimated as 0.14, and the proportion of between-study variance is 0.61. This still is much larger than 0.25, the lower limit for examining differences between studies (Hunter & Schmidt, 1990).

The power of multilevel meta-analysis becomes apparent when we attempt to model the differences in the study outcomes. We simply include the duration of the experimental intervention as an explanatory variable in the model. The multilevel meta-analysis model can be written as

$$d_j = \gamma_0 + \gamma_1 \text{Duration}_{1j} + u_j + e_j \quad (5.6)$$

The advantage of directly including duration as an explanatory variable is that we do not have to dichotomize or discretize it, as we were forced to do in the clustering approach. The results of the multilevel meta-analysis are summarized in Table 5.3. This table presents the results for both the empty (null) model and the model that includes duration, in addition to the results obtained by the classical (random effects) meta-analysis method.

After including duration as an explanatory variable in the model, the residual between-study variance is no longer significant. The regression coefficient for duration is 0.14 ($p < 0.001$), which means that for each additional week, the expected gain in study outcome is 0.14. The explanatory variable duration is centered on its overall mean, and as a

result, the intercept remains essentially unchanged from one model to the next, and it reflects the expected outcome of the average study. The residual variance in the model is 0.04, which is not significant. If we compare this with the between-study variance of 0.14 in the null model, we conclude that 71% of the between-studies variance can be explained by including duration as the explanatory variable in the model.

Because the study outcome depends in part on the duration of the experiment, reporting an overall outcome does not convey all the relevant information. We could report the expected outcomes for different durations (i.e., report the dose-response curve), or calculate which duration is minimally needed to obtain a significant outcome. This is easily done by centering the explanatory variable on different values. For instance, if we center the duration around 2 weeks, the intercept can be interpreted as the expected outcome at 2 weeks. Some multilevel analysis programs can produce predicted values with their expected error variances for various levels of the explanatory variables (such as various durations in this example), which is also useful to describe the expected outcome for experiments with a different duration. Figure 5.1 presents the predicted outcome for our example data for different durations, with the limits of the 95% confidence interval. From the predicted outcome in Fig. 5.1, it is obvious that for low durations negative outcomes are common. Only when the duration of the intervention is at least 4 weeks is the outcome clearly positive.

CORRECTING FOR ARTIFACTS

Hunter and Schmidt (1990, 1994) have advocated to correct study-outcomes for a variety of artifacts. For instance, a common correction is to correct the outcome d 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 (1994) described many other corrections, for instance a correction for the attenuation due to imperfect validity of the outcome measure, corrections for estimated methodological quality, and so on. However, directly applying corrections to the outcome variable results in methodological and statistical problems. A methodological problem is that the majority of these corrections 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. Because these large effects have in fact not been observed,

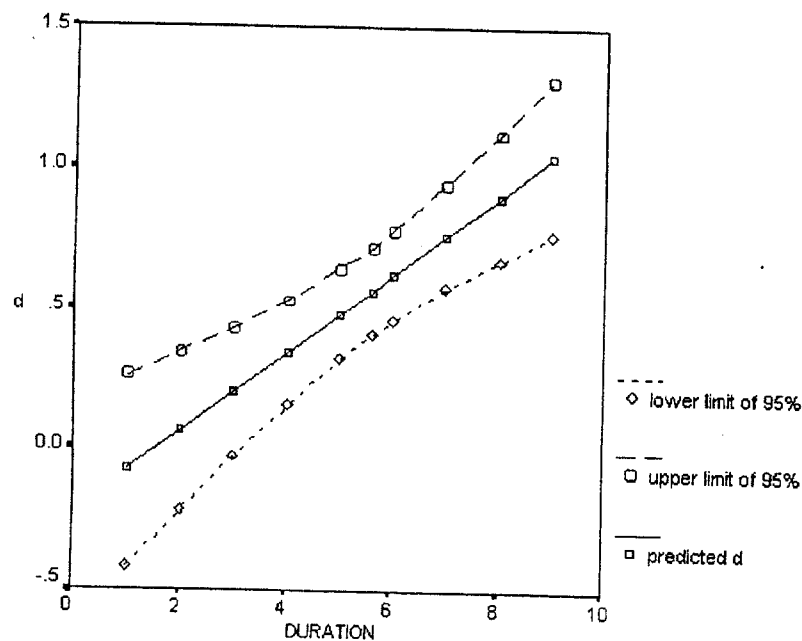


FIG. 5.1. Predicted values for outcome d and 95% interval limits.

automatically carrying out such corrections is controversial. For instance, Schwarzer (1989) advised to always report the original values with the corrected results. A major statistical problem with all these corrections is that their effect on the statistical model is completely unknown. For instance, if the reported reliability is biased, so will be the corrected outcome. If the values used to correct the outcomes are subject to sampling error, and they usually are, the sampling variance of the outcome measure becomes larger. And if many corrections are performed, their cumulative effect on the bias and sampling variance of the outcome measures is totally unclear.

A different and better approach to correct for artifacts is to include them as covariates in the multilevel regression analysis. This is not always optimal; for instance, the attenuation correction follows a multiplicative model, and regression analysis is additive and linear. However, in many cases, adding corrections as explanatory variables to the regression equation produces a reasonable approximation, and when the relationship is not linear, we can always include quadratic or cubic trends in the analysis. For

instance, if the range of reliabilities is not extreme, a linear model for the correction is an acceptable approximation. The advantage of this approach is that the effect of measurement unreliability on the study outcomes is estimated based on the available data instead of on a priori corrections. An additional advantage is that we can test whether the correction has a significant contribution to the regression equation. Lastly, if we suspect that a certain covariate has an effect on the variability of the outcomes, we can include it only in the random part of the model, where it affects the between-study variance, but not the average outcome. This models heteroscedasticity at the study level. For instance, it is reasonable to assume that quality of the experimental design used in a study does not necessarily bias the results, but could result in a larger variability of the outcomes. The result would be a larger variance for the residual errors u_j for studies with a poor experimental design. Of course, some covariates might affect both the average outcome and the study-level variance. Models where the residual error variances are a function of other variables were discussed by Goldstein (1995) under the heading "complex variance structures." Although Goldstein did not discuss their application to multilevel data, this is a straightforward extension of his exposition.

A variation on correcting for artifacts is controlling for the effect of sample 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 larger probability to be 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 is to carry out a fail-safe analysis (Greenhouse & Iyengar, 1994). This answers the question of 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. A different 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 (cf. Light & Pillemer, 1984, Light, Singer & Willet, 1994). If the sample of available studies is "well-behaved" this plot should 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 insignificant small studies remaining unpublished in file drawers.

A problem with the funnel plot is, that we do not know if the smaller studies have different study characteristics too. For instance, small-scale studies could also more often have a short duration. Because the funnel

plot is based on observed outcomes, part of the variability in the plot could be due to explanatory study-level variables. In fact, it would be more appropriate to use a funnel plot after removing the covariate effects. An alternative to the funnel plot is to investigate the effect of the study size directly by including the total sample size of a study as explanatory variable in a multilevel meta-analysis. This allows a formal statistical test, and other study characteristics can be controlled simply by adding these to the explanatory variables.

We illustrate those procedures by correcting our example data for reliability of the measure and for total sample size. Table 5.2 has an entry for reliability (r_{ii}). These fictitious data on the effect of social skill 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 between-study variance is estimated as 0.23 instead of the earlier value of 0.17. The effect of sample size is more difficult to analyze in classical meta-analysis. A funnel plot indicates a well-behaving sample of studies.

The funnel plot shows virtually no relationship between study outcome and total sample size. We can test this more formally by including total sample size as a covariate in the regression model. If we include the sample size and the reliability as explanatory variables in the regression model, we obtain the results presented in Table 5.4.

The first model in Table 5.4 is the empty intercept-only model presented earlier. Model 2 includes the total sample size, centered on its mean of 54.1, as a predictor, and Model 3 the reliability of the outcome measure, centered on the value 1.0, which represents perfect reliability. Model 4 includes the duration of the experiment, centered on its mean of 5.6. Model 5 includes all available predictors. Both the univariate and the simultaneous analysis show that only duration has a significant effect on the study outcome. Differences in measurement reliability and study size pose no major threats to our substantive conclusion about the effect of duration. If reliability is centered on the value 1, the intercept is estimated as 0.67, which is close to the value of 0.64 estimated in the previous section using the classic attenuation correction on the outcomes. However, the large standard error for the reliability slope suggests that this correction is not necessary. Because there is no relation between the study size and the reported outcome, the existence of a file drawer problem related to sample

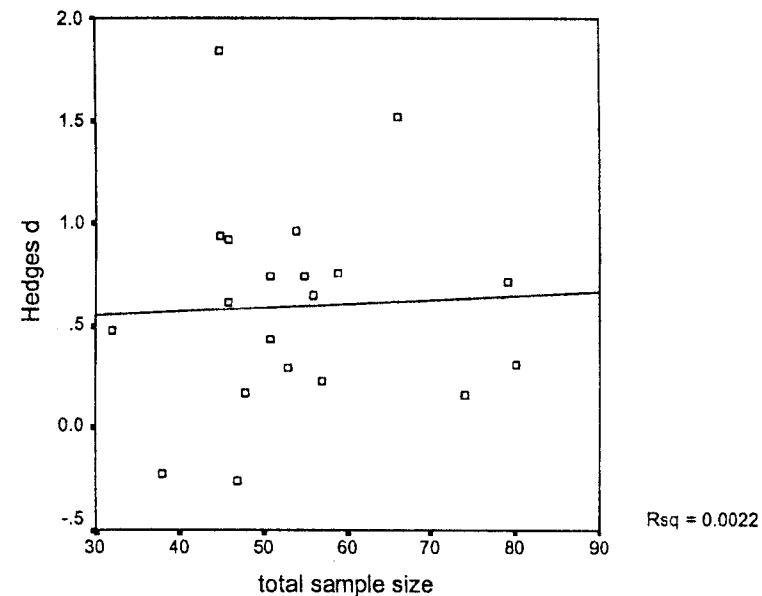


FIG. 5.2. Funnel plot of study outcome against total sample size.

size is unlikely.

The last Model 5 that includes all predictor variables simultaneously is instructive. The (insignificant) regression coefficient for reliability is negative. This is counterintuitive. This is also in the opposite direction of the regression coefficient in Model 3 with reliability as the only predictor. This is a so-called repressor effect caused by the correlations (from 0.25 to 0.33) between the predictor variables. In meta-analysis, because the number of available studies is often small, such effects are likely to occur if we include too many explanatory study-level variables. We conclude that there is an effect of the duration of the treatment on the outcome, and that a bias due to a file drawer problem or differential reliability of the outcome measure is unlikely.

CONCLUSION AND DISCUSSION

The application of multilevel analysis methods in meta-analysis has the advantage that study characteristics can be included in the analysis as potential explanations of the variability of the studies' outcomes. The

TABLE 5.4

Results of Multilevel Regression Including Artifacts as Covariates, all Covariates Centered

Model	1	2	3	4	5
Predictor:	<i>Intercept only</i>	<i>1 plus N_{tot}</i>	<i>1 plus reliab.</i>	<i>1 plus duration</i>	<i>1 plus all predictors</i>
Intercept	.58 (.11)	.58 (.11)	.67 (.28)	.57 (.08)	.49 (.22)
N _{tot}		.001 (.01)			-.004 (.01)
Reliability			.51 (1.48)		-.52 (1.18)
Duration				.14 (.03)	.15 (.04)
Parameters					
Variance					
σ_u^2	.14	.16	.16	.04	.05
χ^2 test					
p value	$p < .001$	$p < .001$	$p < .001$	$p = .09$	$p = .07$

study characteristics can be theoretically important constructs, or they can be covariates intended to correct for possible artifacts. Significance tests of the regression coefficients and predictions can be used to assess the effect of study characteristics. The residual study-level variance can be tested for significance to assess whether the study variables explain all the between-study variance. A comparison of the variance in the empty (null) model and in the final model informs us how much variance between study outcomes is explained by our model.

Multilevel regression analysis assumes that all relations are additive and linear. An additional assumption is that the distribution of the outcomes is normal (most classical meta-analysis methods require the same assumption). If these assumptions are violated, transforming the outcome measure can be helpful (cf. Bryk & Raudenbush, 1992; Hedges & Olkin, 1985). In practice, we usually have unexplained between-study variance, as violations of assumptions in the original analyses that produced the published outcome measures tend to lead to increased variability of the outcomes. For this reason, we should preferably use models that include between-studies variance, such as random effects meta-analysis and multilevel models. Hunter and Schmidt (1994) considered a between-study variance of up to 25% of the total variance to be uninteresting. In their view, a larger amount of between study variance cannot be attributed to various artifacts, and should be further investigated. Following this line of reasoning, we argue that when we add explanatory variables to a model,

obtaining a residual variance of less than 25% of the total variance is a sign that our model is reasonably complete.

An interesting extension of the multilevel regression model discussed here is allowing for more than two levels. For instance, we may have a situation where there are several outcome measures for each study. The approach in classical meta-analysis is to either combine these into one single outcome per study, or to carry out separate meta-analyses for each different outcome (Gleser & Olkin, 1994). In a multilevel model, it is possible to specify a multivariate outcome model. When all studies report all available outcome measures, the multivariate multilevel model is a straightforward extension of the univariate model (cf. Raudenbush & Bryk, 1985). When some studies do not report on all available outcomes, we have a missing data problem. This extension leads to a more complicated model, which still can be estimated using standard multilevel software. For details, see Kalaian and Raudenbush (1996) and Goldstein (1995). A related extension arises when we have summary data for some studies, whereas we have access to the raw data for others. Goldstein and Yang (2000) showed that such data can be combined in a single model, using standard multilevel analysis software. This allows more refined analyses, using all available data.

The program HLM (Bryk, Raudenbush & Congdon, 1994; Raudenbush, Bryk, Cheong & Congdon, 2000) has a built-in provision for meta-analysis that is restricted to two-levels. If we need three levels, we can use the standard HLM/3L software, using an adapted program setup. The software MLn/MLwiN (Rasbash & Woodhouse, 1995) can also be used for meta-analysis, again with an adapted setup. Ways of tweaking standard multilevel software for meta-analysis are discussed in the Appendix.

There are some minor differences between the programs. HLM uses by default an estimator based on restricted maximum likelihood (RML), whereas MLwiN by default uses full maximum likelihood (FML, called IGLS in MLwiN). Because 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). The results reported earlier were computed using RML. If FML is used, the differences turn out to be small.

An important difference between HLM and MLwiN is the test used to assess the significance of the variances. HLM by default uses a variance test based on a chi-square test of the residuals (Bryk & Raudenbush, 1992). 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. It is based on the assumption of normality; variances have a chi-square distribution. Especially with small sample sizes and small variances, the Z-test may be inaccurate. An additional advantage of the

chi-square test on the residuals is that for the null model it is equivalent to the chi-square variance test in classical meta-analysis (Hedges & Olkin, 1985). The variance tests reported earlier used the chi-square test on the residuals. MLwiN does not offer this test, but it can be produced using the MLwiN macro language.

For estimating complex models, Bayesian procedures are promising and coming into use. These use computer-intensive methods such as Markov Chain Monte Carlo (MCMC) methods to estimate the parameters and their sampling distributions. These methods are attractive for meta-analysis (DuMouchel 1994), because they are less sensitive to the problems that arise when we model small variances in small samples. Bayesian models can be extended by including a prior distribution. This prior distribution reflects a priori beliefs about the likelihood of publication bias. In principle, this is an elegant method to investigate the effect of publication bias. An example of such an analysis is found in Tweedie, Scott, Biggerstaff, and Mengersen (1994). Present multilevel software cannot analyze such models, and more complicated software is needed, such as the general Bayesian modeling program BUGS (Spiegelhalter, 1994).

APPENDIX

Software issues

The simplest program for multilevel meta-analysis is VKHLM, which comes with HLM 2.0 (Bryk et al., 1994) as a separate program, and is built into HLM as an option in later versions. HLM expects for each study a row of data containing a study ID, an outcome measure, its sampling variance, followed by the explanatory variables. If the null model is specified, the results from HLM are close to the classical meta-analysis results produced by Schwarzer's program META, provided one realizes that META reads effect sizes g and transforms these automatically into d 's.

Using MLwiN or 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 sampling error is included only in the random part on level 1, with a coefficient fixed at 1 (MLwiN uses the command RCON for this). 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

$$\chi^2 = \sum \left[\left(d_j - \hat{d}_j \right) / SE(d_j) \right]^2,$$

the sum of the squared residuals divided by their sampling variances. The degrees of freedom are given by $df = J - p - 1$, where J is the number of studies, and p the number of explanatory variables in the model. Assuming that the outcomes are denoted by d , and the standard errors by sed , the sequence of MLwiN commands for computing the chi-square is: PRED C50; CALC C50=[(d-C50)/sed]^2; SUM c50 to B1; CPRO B1 df. This code assumes that the spreadsheet column C50 is unused.

If we need more than two levels in HLM, we must use HLM/3L, which does not include the VKHLM option. For HLM/3L, we also need a special setup. In this case, we include the standard errors as a weighting variable at the lowest level. We must instruct the program *not* to normalize the weights, which is the default option, and constrain the lowest level variance to be equal to 1.

To apply multilevel models in meta-analysis in other software, this 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. These options are available in the multilevel options in Lisrel 8.3 (du Toit, du Toit, Jöreskog, & Sörbom, 1999) and in the multilevel analysis program aML (Lillard & Panis, 2000), but these programs do not include the recommended RML estimation. So far, commonly available public domain software for multilevel analysis, such as MixReg (Hedeker & Gibbons, 1996) does not offer the necessary options.

For classical meta-analysis, the program META (Schwarzer, 1989) is freely available from the Internet location <http://userpage.fu-berlin.de/health/meta.htm>. It comes with a program manual that also explains the basic elements of meta-analysis. The program MetaWin (Rosenberg, Adams & Gurevitch (2000) contains a limited weighted least squares regression option. As indicated earlier, the iterated maximum likelihood methods employed in multilevel analysis is generally more efficient.

References

- Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

- Bryk, A. S., Raudenbush, S. W. & Congdon, R. T. (1994). *HLM 2/3. Hierarchical linear modeling with the HLM/2L and HLM/3L programs*. Chicago: Scientific Software International.
- Cooper, H. (1998). *Synthesizing research: A guide for literature reviews*. Thousand Oaks, CA: Sage.
- Cornell, J., & Mulrow, C. (1999). Meta-analysis. In H. J. Ader & G. J. Mellenbergh (Eds.), *Research methodology in the social, behavioral, and life sciences*, (pp. 285-323). London: Sage.
- DuMouchel, W. H. (1994). *Hierarchical Bayesian linear models for meta-analysis*. Washington, DC: National Institute of Statistical Sciences.
- du Toit, S., du Toit, M., Jöreskog, K. G., & Sörbom, D. (1999). *Interactive LISREL user's guide*. Chicago: Scientific Software Inc.
- Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 10, 3-8.
- Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-356). New York: Russel Sage Foundation.
- Greenhouse, J. B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 383-398). New York: Russel Sage Foundation.
- Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.
- Goldstein, H., & Yang, M. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics*, 49 (3), 399-412.
- Hedeker, D., & Gibbons, R. D. (1996). MIXREG: A computer program for mixed effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine*, 49, 229-252.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects in meta-analysis. *Psychological Methods*, 3 (4), 486-504.
- Hox, J. J., 1995. *Applied multilevel analysis (2nd ed.)*. Amsterdam: TT-Publikaties.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis* (pp. 323-336). Newbury Park, CA: Sage.
- Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artifact variation. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis*. New York: Russel Sage Foundation.
- Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, 1, 227-235.

- Light, R. J. & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Light, R. D., Singer, J. D., & Willet, J. B. (1994). The visual presentation and interpretation of meta-analyses. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 439-454). New York: Russell Sage Foundation.
- Lillard, L. A., & Panis, C. W. A. (2000). *aML. Multilevel multiprocess statistical software, (release 1)*. Los Angeles, CA: EconWare.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.
- Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator effects. *Psychological Methods*, 3 (3), 354-379.
- Rasbash, J., & Woodhouse, G. (1995). *MLn command guide*. London: Multilevel Models Project, University of London.
- Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-322). New York: Russell Sage Foundation.
- Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, 10, 75-98.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2000). *HLM5. Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software Inc.
- Rosenberg, M. S., Adams, D. C., & Gurevitch, J. (2000). *MetaWin. Statistical software for meta-analysis*. Sunderland, MA: Sinauer Associates.
- Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.
- Rosental, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.
- Schwarzer, R. (1989). *Meta-analysis programs [computer program manual]*. Berlin: Institut für Psychologie, Freie Universität Berlin.
- Snijders, T. A. B., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.
- Spiegelhalter, D. (1994). *BUGS: Bayesian inference using Gibbs sampling*. Cambridge, England: MRC Biostatistics Unit.
- Tweedie, R. L., Scott, D. J., Biggerstaff, B. J., & Mengersen, K. L. (1994). *Bayesian meta-analysis, with application to studies of ETS and lung cancer*. Unpublished report, Colorado State University, Fort Collins.