31

Analysis of Moderator Effects in Meta-Analysis

Wolfgang Viechtbauer

As discussed in Chapter 12 (this volume), meta-analysis is a quantitative method for leveraging the proliferation of published research to more scientifically and comprehensively synthesize bodies of research (e.g., Chalmers, Hedges, & Cooper, 2002). Social policy decisions and best practices in various fields are increasingly influenced not by the results from single isolated studies but by the findings from meta-analyses (Cook et al., 1992).

A meta-analysis not only helps to determine whether a particular treatment is actually effective or whether there is indeed an association between variables but also allows the reviewer to examine whether the treatment effectiveness or relationship strength is influenced by the characteristics of the studies. For example, it is conceivable that the effectiveness of a treatment observed in a particular study depends on the treatment duration or intensity (e.g., the length of the psychotherapy or the medication dosage), the characteristics of the sample, the study setting, or the type of outcome measure used. Examining these hypotheses is difficult when conducting a traditional narrative literature review, but such *moderator analyses* constitute an integral and important aspect of a meta-analysis (Lau, Ioannidis, & Schmid, 1998; Thompson, 1994).

Scope of This Chapter

One can roughly break the process of a meta-analysis down into five stages (Cooper, 1998): (a) problem formulation, (b) data collection, (c) data evaluation, (d) analysis and interpretation, and (e) presentation of results. The majority of the time and effort will typically be spent on the first three stages, which are briefly outlined in Chapter 12 and elsewhere.[1] For the purposes this chapter, we will assume that these steps have already been completed. Instead, the present chapter is meant to provide some guidelines on how to conduct the statistical analysis, once the first three steps have been completed. Again, several books deal extensively with this topic (e.g., Cooper & Hedges, 1994; Hedges & Olkin, 1985), and a single chapter cannot replace these references. However, the statistical methods that should be used for a meta-analysis are constantly being improved and extended. The goal is then to highlight those methods that currently represent best practices.

A Sample Data Set

An example will be used throughout this chapter to make the discussion more concrete and to allow the reader to experiment with the

techniques discussed. Consider Table 31.1, which provides the results from $k = 16$ studies examining the effectiveness of massage therapy for reducing state anxiety. In each study, the amount of anxiety was measured among subjects randomly assigned to either a massage therapy or a control/standard treatment group. For each study (i being the index for the studies), the table lists the sample size of the control/comparison and the treatment group (n_i^C and n_i^E, respectively), the effect size estimate (Y_i) in the form of a standardized mean difference (to be discussed in more detail below), the estimated sampling variance ($\hat{v}_i$) of the effect size estimate, the minutes per session of massage therapy provided, whether a fully trained massage therapist or layperson provided the therapy (coded as 1 and 0, respectively), the mean age of the sample, and whether the study was conducted by the Touch Research Institute (TRI) or not (coded as 1 and 0, respectively). The last four variables are examples of *moderator variables* that may influence the effectiveness of massage therapy for reducing state anxiety.[2]

Standardized Mean Difference

The standardized mean difference (SMD) is usually the effect size measure of choice when we are interested in the difference between a treatment/experimental and a control/comparison group and the outcome variable is quantitative. Let μ_i^C and μ_i^E denote the true (population) means of the control and treatment groups in the ith study, and let σ_i denote the common standard deviation of these groups. Then the effect size in the ith study is defined as

$$\theta_i = \frac{\mu_i^C - \mu_i^E}{\sigma_i},$$

so that θ_i indicates, in standard deviation units, by how much the mean in the control group differs

Table 31.1 Results From 16 Studies on the Effectiveness of Massage Therapy for Reducing State Anxiety

	Sample Size							
Study	n_i^C	n_i^E	*Effect Size Estimate* (Y_i)	*Sampling Variance* ($\hat{v}_i$)	*Minutes per Session*	*Trained Therapist*	*Mean Age*	*TRI Study*
1	30	30	0.444	0.068	30	0	28	0
2	39	46	–0.495	0.049	10	0	42	0
3	15	15	0.195	0.134	20	1	31	0
4	10	10	0.546	0.207	40	1	39	1
5	12	12	0.840	0.181	20	1	17	1
6	10	10	0.105	0.200	30	1	51	1
7	24	26	0.472	0.082	15	1	26	1
8	14	18	–0.205	0.128	10	0	64	0
9	12	12	1.284	0.201	45	1	48	1
10	12	12	0.068	0.167	30	1	40	1
11	15	15	0.234	0.134	30	1	52	1
12	12	12	0.811	0.180	30	1	33	1
13	15	15	0.204	0.134	30	1	20	1
14	18	18	1.271	0.134	60	1	27	0
15	15	15	1.090	0.153	45	1	52	0
16	35	43	–0.059	0.052	10	1	61	0

NOTE: Adapted from Moyer, Rounds, and Hannum (2004), leaving out three studies with missing data and two studies where the duration of massage therapy provided was less than 10 minutes. Minutes per session = minutes of therapy provided per session; trained therapist = 0 for a layperson providing the therapy and 1 for a trained therapist; mean age = mean age of the sample; TRI study = 1 when the study was conducted by the Touch Research Institute (TRI) and 0 otherwise.

from that of the experimental group after the treatment. For the massage therapy meta-analysis, θ_i can be interpreted as a measure of the effectiveness of massage therapy for reducing state anxiety. Specifically, positive values of θ_i indicate lower amounts of anxiety in the treatment group, values around zero indicate no difference between the two groups, and negative values indicate lower anxiety in the control group. Expressing the effect size in standard deviation units makes the results from studies using different outcome measures (i.e., studies using anxiety scales with different raw units) comparable.

Hedges (1981) showed that an approximately unbiased and normally distributed estimate of θ_i is given by

$$Y_i = \left(1 - \frac{3}{4(n_i^C + n_i^E) - 9}\right)\left(\frac{\bar{x}_i^C - \bar{x}_i^E}{s_i}\right), \quad (1)$$

where $\bar{x}_i^C$ and $\bar{x}_i^E$ are the observed means of the treatment and control group in the ith study, and s_i is the pooled standard deviation of the two groups. The sampling variance of Y_i can be estimated with

$$\hat{v}_i = \frac{n_i^C + n_i^E}{n_i^C n_i^E} + \frac{Y_i^2}{2(n_i^C + n_i^E)}. \quad (2)$$

Therefore, Y_i is an estimate of θ_i, and $\hat{v}_i$ is an estimate of the amount of variability in Y_i we would expect due to subject-level sampling variability. In other words, even if the θ_i values (i.e., the true SMDs) are identical in two studies, we would not expect the corresponding Y_i values (i.e., the observed SMDs) to coincide due to sampling differences among the samples. However, should the sample sizes be very large in the two studies, then $\hat{v}_i$ is small (which should be evident from Equation 2); hence, sampling variability decreases, and the two Y_i values would tend to be very close to each other.

Meta-Analytic Data in General

The SMD is not the only effect size measure used in meta-analyses. Others include correlation coefficients and odds ratios (as discussed in Chapter 17).[3]

However, regardless of the specific effect size measure used in a meta-analysis, assume that k independent effect size estimates have been collected along with information about one or more moderator variables. As discussed earlier, each effect size estimate Y_i is an estimate of a corresponding parameter θ_i, which indicates the true effect size in the ith study. In general, we can express this idea by writing

$$Y_i = \theta_i + \varepsilon_i, \quad (3)$$

where ε_i is the sampling error for the ith study. The sampling errors are assumed to be normally distributed with mean zero and variance $\hat{v}_i$.

Meta-Analytic Models

Once a collection of effect size estimates (like the one in Table 31.1) has been obtained, several questions arise:

1. Is massage therapy an effective treatment for reducing state anxiety (i.e., how large is the overall effect of massage therapy on state anxiety?)?
2. Does the treatment effectiveness vary across studies and, if yes, by how much (i.e., is the effect size the same in all studies, and if not, how much variability is there among the effect sizes?)?
3. If there is variability in the treatment effectiveness across studies, is this variability, at least in part, systematic and explainable (i.e., do the effect sizes depend on one or more moderators—in particular, the treatment duration, the level of training of the therapist, the mean age of the sample, or whether the study was conducted by the TRI or not?)?

To answer these questions, we must identify the model that most closely approximates the true structure underlying the collection of effect size estimates.

Fixed Effects Model

The simplest case we may consider is the fixed effects model. According to this model, the effect sizes are homogeneous (i.e., $\theta_1 = \ldots = \theta_k$), so the model is given by

$$Y_i = \theta + \varepsilon_i, \quad (4)$$

where θ denotes the (homogeneous) effect size for all k studies. In the context of the massage therapy

meta-analysis, this would imply that the treatment effectiveness is the same in all studies, regardless of treatment duration, level of training of the therapist, mean age of the sample, whether the study was conducted by the TRI or not, or any other moderator variable that we did not collect any information on (i.e., the studies may differ in other aspects unknown to us). Therefore, differences among the observed standardized mean differences (i.e., the effect size estimates) are assumed to be a result of sampling variability alone.

Fixed Effects With Moderators Model

On the other hand, when the effect sizes are not all equal to each other, they are said to be heterogeneous. Heterogeneity among the effect sizes may be a result of moderators and therefore entirely systematic. For example, when the effectiveness of massage therapy increases with the minutes of treatment provided and/or the training level of the therapist, then θ_i will be systematically higher in studies where the duration of the therapy was longer and/or the therapy was provided by a trained massage therapist as opposed to a layperson. Differences between the effect size estimates are then not only a result of sampling variability but also a result of the influence of moderators on the effect sizes. This case can be described by a fixed effects with moderators model, which is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i, \quad (5)$$

where X_{ij} denotes the value of the jth moderator variable for the ith effect size. The fixed effects with moderators model therefore assumes that the effect sizes are a linear function of one or more moderator variables. For example, in the massage therapy meta-analysis, the true effectiveness of the treatment may be a linear function of the $p = 4$ moderator variables described earlier.

Random Effects Model

Alternatively, the heterogeneity among the effect sizes may be completely random (unsystematic). In that case, the θ_i values will differ from each other randomly, and it will not be possible to account for differences among the effect sizes based on moderator variables such as treatment duration or the training level of the therapist. In this case, the random effects model applies, which is given by

$$Y_i = \mu + u_i + \varepsilon_i, \quad (6)$$

where μ denotes the average effect size, and u_i is assumed to follow a normal distribution with mean zero and variance τ^2_{RE}. Therefore, τ^2_{RE} denotes the total amount of heterogeneity among the effect sizes. Differences between the effect size estimates are now assumed to be a result of sampling variability and random differences among the effect sizes.

Mixed Effects Model

Finally, it is possible that the heterogeneity among the effect sizes is, in part, a result of moderators and, in part, random. In that case, the mixed effects model applies, which is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + u_i + \varepsilon_i. \quad (7)$$

The variance of u_i, now denoted by τ^2_{ME}, represents the amount of *residual heterogeneity,* that is, the amount of excess or unexplainable variability in the effect sizes (i.e., heterogeneity that cannot be accounted for by the moderator variables included in the model). Therefore, the mixed effects model assumes that the effect sizes are a linear function of one or more moderator variables but also allows for the possibility that residual heterogeneity may exist in the effect sizes. It is therefore the most general of the four meta-analytic models and should be the starting point in most meta-analyses (this point will be elaborated on below). In fact, it is easy to see that the fixed effects, fixed effects with moderators, and random effects models are just special cases of the mixed effects model.[4] The nested hierarchy among the four models is illustrated in Figure 31.1. For further discussion of these models, see, for example, Hedges (1994), Hedges and Olkin (1985), and Raudenbush (1994).

Note that the mixed effects and fixed effects with moderators models can accommodate quantitative and categorical moderator variables. For categorical moderator variables, one has to employ an appropriate coding scheme as used in regression analysis when including categorical independent variables in the model (e.g., Neter, Kutner, Nachtsheim, & Wasserman, 1996).

Why Start With the Mixed Effects Model?

The mixed effects model was earlier suggested as the starting point for meta-analyses that are

focusing on moderators. This recommendation actually goes contrary to typical practice, as meta-analysts usually first report an overall effect size estimate from a fixed or random effects model before considering the influence moderators. However, several reasons speak against this practice. First of all, heterogeneity is typically present among the effect sizes. Empirical evidence strongly suggests that the effect sizes are influenced considerably, for example, by the methods and procedures used in the studies, the characteristics of the samples, the study settings, or the types of outcome measures used (e.g., Wilson & Lipsey, 2001). Exploring the source of the heterogeneity by examining the influence of moderators on the effect sizes is often one of the most important and useful aspects of a meta-analysis (Lau et al., 1998; Thompson, 1994).

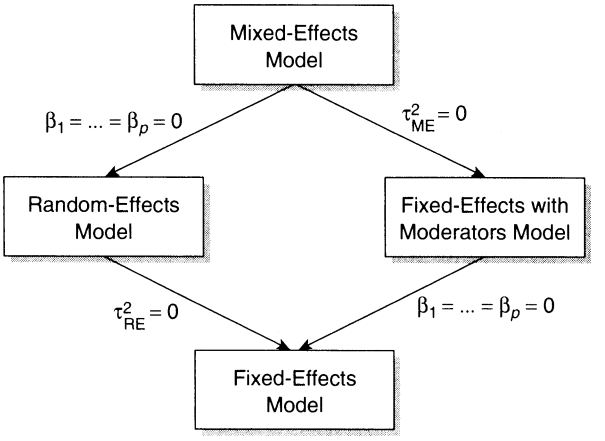

Figure 31.1 The nested hierarchical structure between the meta-analytic models.

In fact, an estimate of the overall effect size is meaningless at best and can even be misleading when moderators are present. Consider, for example, the admittedly extreme but illustrative case where (a) the effectiveness of massage therapy depends only on whether the treatment is given by a trained therapist or by a layperson, (b) the true SMD is equal to 0.5 in studies using a trained therapist and equal to –0.5 in studies using a layperson (i.e., massage therapy given by a trained therapist results in decreased anxiety levels, while a layperson does more harm than good and actually causes an increase in anxiety levels), and (c) the treatment was given by a trained therapist in about half of the studies, while the other half used a layperson. Then an estimate of the overall effect size would fall around zero, suggesting the total absence of an effect.

Moreover, it is unclear what such an overall effect size estimate represents. For the dichotomous moderator that distinguishes between a layperson and a trained therapist, an overall estimate may, with some imagination, represent the effect size for a semi-trained therapist. However, for the dichotomous moderator that distinguishes between studies conducted by the TRI and other laboratories, it is difficult to imagine what such an estimate would describe. Therefore, in those cases where moderators influence the effect sizes, one should resist the temptation to oversimplify matters by reporting a single overall effect size estimate. This implies that we should actually start out with a model that examines the influence of moderators on the effect sizes. If the influence of the moderators is large (either practically speaking or in terms of statistical significance), then one can provide estimated or predicted effect sizes for some sensible values of the moderator variables based on the fitted model (to be illustrated later on).

Another issue to consider in this context is the use of models that acknowledge the possible presence of (residual) heterogeneity. Although the issue continues to be debated in the literature, a general consensus is beginning to emerge that one should employ random/mixed instead of fixed effects procedures, at least as a starting point in the analysis (e.g., Field, 2003; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; National Research Council, 1992; Overton, 1998). Given the hierarchical nature of the models (cf. Figure 31.1), the mixed or random effects models may ultimately reduce to a simpler model. Specifically, when residual heterogeneity is absent (i.e., $\tau^2_{ME} = 0$), the mixed effects model automatically simplifies to a fixed effects with moderators model, while the random effects model automatically simplifies to a fixed effects model when there is no heterogeneity at all (i.e., $\tau^2_{RE} = 0$). Therefore, instead of adopting a simpler (and possibly incorrect) model a priori, we should examine what model is actually supported by the data.

Fitting the Mixed Effects Model

Fitting the mixed effects model is done in two steps. First, we estimate τ^2_{ME}, the amount of

residual heterogeneity in the effect sizes. We then estimate β_0 through β_p, the parameters specifying the relationship between the effect sizes and the moderators.

Estimating the Amount of Residual Heterogeneity

Numerous methods for estimating τ^2_{ME} have been discussed in the literature (e.g., Raudenbush, 1994; Raudenbush & Bryk, 1985; Sidik & Jonkman, 2005; Thompson & Sharp, 1999), but a description of the various methods is beyond the scope of the present chapter. Here, we will simply focus on a commonly used method of moments estimator (e.g., Raudenbush, 1994).

Let the $(k \times (p + 1))$ matrix X contain the values of the p moderator variables to be included in the model, where the first column consists of a vector of 1s, corresponding to the intercept parameter β_0. Also, collect the effect size estimates into the $(k \times 1)$ vector y. Next, let $w_i = 1/\hat{v}_i$, and define W as the diagonal matrix using those weights. Now calculate

$$\mathbf{P} = \mathbf{W} - \mathbf{WX}\,(\mathbf{X'WX})^{-1}\,\mathbf{X'W}$$

and finally

$$\hat{\tau}^2_{ME} = \frac{\mathbf{y'Py} - (k - p - 1)}{tr[\mathbf{P}]}, \qquad (8)$$

where $\mathbf{X'}$ denotes the transpose of X and $\mathbf{y'}$ the transpose of y, $(\mathbf{X'WX})^{-1}$ denotes the inverse of $(\mathbf{X'WX})$, and $tr[\mathbf{P}]$ denotes the trace of the P matrix. Should the estimate be negative, then this indicates the absence of residual heterogeneity, and we set $\hat{\tau}^2_{ME} = 0$.

Illustrative Example

Four moderators will be included in the model for the massage therapy meta-analysis—namely, the minutes of therapy provided, whether a layperson or a fully trained massage therapist provided the therapy, the mean age of the sample, and whether the study was conducted by the TRI or some other laboratory. Therefore, X, y, and W are given by

$$\mathbf{X} = \begin{bmatrix} 1 & 30 & 0 & 28 & 0 \\ 1 & 10 & 0 & 42 & 0 \\ 1 & 20 & 1 & 31 & 0 \\ 1 & 40 & 1 & 39 & 1 \\ & & \vdots & & \\ 1 & 10 & 1 & 61 & 0 \end{bmatrix},$$

$$\mathbf{y} = \begin{bmatrix} 0.444 \\ -0.495 \\ 0.195 \\ 0.546 \\ \vdots \\ -0.059 \end{bmatrix}, \text{ and}$$

$$\mathbf{W} = \begin{bmatrix} \frac{1}{0.068} & 0 & 0 & 0 \\ 0 & \frac{1}{0.049} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{1}{0.052} \end{bmatrix}.$$

Applying Equation 8 then yields $\hat{\tau}^2_{ME} = -0.023$. Since the estimate is negative, we set $\hat{\tau}^2_{ME} = 0$ and conclude that no residual heterogeneity is present, or, in other words, the moderators included in the model account for all of the heterogeneity in the effect sizes.

Estimating the Moderator Parameters

Having obtained an estimate of τ^2_{ME} (with this or any other method), we can then estimate β_0 through β_p with

$$\mathbf{b} = (\mathbf{X'WX})^{-1}\,\mathbf{X'}\,\mathbf{Wy}, \qquad (9)$$

where the elements of the diagonal W matrix are now set to $w_i = 1/(\hat{\tau}^2_{ME} + \hat{v}_i)$. The variance-covariance matrix of the parameter estimates in b is then obtained with

$$\hat{\Sigma} = (\mathbf{X'WX})^{-1}. \qquad (10)$$

Taking the square root of the diagonal elements of $\hat{\Sigma}$ yields the standard errors of the estimates, which will be denoted by $SE[b_j]$.

Illustrative Example

Since the estimate of the amount of residual heterogeneity happened to be zero in our example, $w_i = 1/(\hat{\tau}^2_{ME} + \hat{v}_i)$ actually simplifies to $w_i = 1/\hat{v}_i$, and the **W** matrix remains unchanged. The parameter estimates and variance-covariance matrix obtained by applying Equations 9 and 10 are equal to

$$\mathbf{b} = \begin{bmatrix} -0.263 \\ 0.025 \\ 0.338 \\ -0.007 \\ -0.061 \end{bmatrix} \text{ and } \hat{\Sigma} = \begin{bmatrix} .13555 & -.00139 & .00457 & -.00215 & -.02604 \\ -.00139 & .00005 & -.00055 & .00002 & .00013 \\ .00457 & -.00055 & .05620 & -.00045 & -.02970 \\ -.00215 & .00002 & -.00045 & .00005 & .00058 \\ -.02604 & .00013 & -.02970 & .00058 & .04885 \end{bmatrix}.$$

Therefore, for a 1-minute increase in session duration, the effectiveness of massage therapy is estimated to increase by $b_1 = 0.025$ points in SMD units ($SE[b_1] = \sqrt{.00005} = .007$). For example, an increase in 12 minutes should result in a 0.3 increase in the effect size. Lacking further information about the domain being studied, 0.2, 0.5, and 0.8 are conventionally thought of as small, medium, and large SMDs (Cohen, 1988). Therefore, 12 minutes can mean the difference between a small and a medium or a medium and a large effect. Moreover, the effect size is estimated to be $b_2 = 0.338$ points higher for a trained massage therapist when compared with a layperson providing the treatment ($SE[b_2] = \sqrt{.05620} = .237$). Furthermore, for a 1-year increase in the average age of the sample, the effect size is estimated to change by $b_3 = -0.007$ points ($SE[b_3] = \sqrt{.00005} = .007$).

Finally, studies conducted by the TRI are estimated to yield an effect size that differs by $b_4 = -0.061$ SMD units from that of other laboratories ($SE[b_4] = \sqrt{.04885} = .221$). The b_0 value should not be interpreted here, as it estimates the effectiveness of zero minutes of therapy provided by a layperson to a sample with an average age of zero in a study that was conducted in a laboratory other than the TRI.[5]

Returning to the point made earlier about avoiding a single overall effect size in the presence of moderators, we may now report the estimated effect size for some sensible and representative moderator values. For example, the estimated effect size for 10 minutes of massage therapy provided by a layperson in a non-TRI study to a group with an average age of 40 is –0.293 (i.e., –0.263 + 0.025(10) + 0.338(0) – 0.007(40) – 0.061(0) = –0.293). On the other hand, 30 minutes of therapy provided by a trained therapist to the same group in a non-TRI study is estimated to yield an effect size of 0.545 (i.e., –0.263 + 0.025(30) + 0.338(1) – 0.007(40) – 0.061(0) = 0.545). The estimated effect is actually negative in the first and positive in the second case, and the difference between the two amounts to more than 0.8 SMD units. A simple average would not be able to properly represent such differences.

Moderator Analysis

Although we have already seen that the estimated effect changes drastically as a function of the moderators, we may want to test whether the moderators included in the model exert a statistically significant influence on the effect sizes in general. Also, when several moderators are included in the model, we may want to examine the statistical significance of each moderator variable individually. Refined procedures for carrying out such tests, which have been developed in recent years (e.g., Knapp & Hartung, 2003; Sidik & Jonkman, 2003, 2005), will be discussed in the present section.

We start by calculating an adjusted variance-covariance matrix with

$$\hat{\Sigma}^* = s^2_{\hat{w}} (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}, \tag{11}$$

where either

$$s^2_{\hat{w}} = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b})' \mathbf{W} (\mathbf{y} - \mathbf{X}\mathbf{b})}{k - p - 1} \tag{12}$$

do not truncate to 1 (just leave (12) as is, even if it below 1).

or it is set equal to 1 if the value calculated with Equation 12 falls below 1.

Omnibus Test of all Moderators

When multiple moderators are included in the model, we can test the null hypothesis H_0: $\beta_1 = \ldots = \beta_p = 0$ (i.e., whether *any* of the moderator variables are related to the effect sizes) by computing

$$Q_R = \mathbf{b}'_{[2]} (\hat{\Sigma}^*_{[2]})^{-1} \mathbf{b}_{[2]}, \qquad (13)$$

where $\mathbf{b}_{[2]}$ is the $(p \times 1)$ vector of parameter estimates excluding the first element (which corresponds to the intercept estimate, which we do not want to include in the test), and $\hat{\Sigma}^*_{[2]}$ is the lower right $(p \times p)$ matrix obtained from $\hat{\Sigma}^*$ after deleting the first column and first row. We compare the Q_R value against $p \times F(\alpha; p, k-p-1)$, where $F(\alpha; df_1, df_2)$ denotes the critical value of an F distribution with df_1 and df_2 degrees of freedom at the desired α–level. If $Q_R > p \times F(\alpha; p, k-p-1)$, we reject the null hypothesis and conclude that at least one of the moderators is related to the effect sizes. Otherwise, we conclude that the effect sizes are not influenced by any of the moderators included in the set that was tested.

Illustrative Example

do not truncate to 1 (this is overly conservative)

Applying Equation 12 after we have fitted the mixed effects model yields a value of $s^2_{\dot{w}} = 0.822$. Therefore, $s^2_{\dot{w}}$ is set equal to 1, so that the adjusted variance-covariance matrix, obtained with Equation 11, is identical to the one given earlier. Finally, to test whether at least one moderator is related to the effect sizes, we apply the Q_R test (Equation 13), with

$$\mathbf{b}_{[2]} = \begin{bmatrix} 0.025 \\ 0.338 \\ -0.007 \\ -0.061 \end{bmatrix} \text{ and } \hat{\Sigma}^*_{[2]} =$$

$$\begin{bmatrix} .00005 & -.00055 & .00002 & .00013 \\ -.00055 & .05620 & -.00045 & -.02970 \\ .00002 & -.00045 & .00005 & .00058 \\ .00013 & -.02970 & .00058 & .04885 \end{bmatrix}.$$

The value of Q_R is 28.91, which we compare against $4 \times F(.05; 4, 11) = 13.43$. We therefore conclude that at least one of the moderator variables influences the effectiveness of massage therapy.

Individual Moderator Tests

We can also test the statistical significance of each moderator variable individually with

$$t_{b_j} = \frac{b_j}{s_{\dot{w}} SE[b_j]}, \qquad (14)$$

which we compare against the critical values of a t distribution with $k-p-1$ degrees of freedom. Alternatively,

$$b_j \pm t_{(k-p-1;\ 1-\alpha/2)}\ s_{\dot{w}}\ SE[b_j] \qquad (15)$$

provides a $(1-\alpha) \times 100\%$ confidence interval for β_j.

Illustrative Example

Since $s^2_{\dot{w}}$ was set equal to 1, the t_{b_j} values (Equation 14) are obtained by dividing the parameter estimates by the standard errors given earlier. We compare these values against $\pm$ 2.20, the critical bounds of a t-distributed random variable with 16 − 4 − 1 = 11 degrees of freedom (using α = .05, two-tailed). Alternatively, 95% confidence intervals can be computed with Equation 15. These results are summarized in Table 31.2, which indicates that only the minutes per session moderator is statistically significant.

The effect size estimates are shown in Figure 31.2 after ordering the studies by the minutes per session moderator variable. The approximate bounds of individual 95% confidence intervals are also shown, which are given by

$$Y_i \pm 1.96\sqrt{\hat{\tau}^2_{ME} + \hat{v}_i}$$

(note that τ^2_{ME} happens to be zero in this particular case). As Figure 31.2 clearly demonstrates, the effect size estimates tend to increase systematically with treatment duration. Such a pattern

Table 31.2 Results From Fitting the Mixed Effects Model to the Data in Table 31.1 When Entering All Moderators Simultaneously in the Model

Moderator	b_j	$s_{\dot{w}}\, SE\,[b_j]$	t_{b_j}	*95% CI for* β_j
Intercept	–0.263	0.368	–0.71	(–1.07, 0.55)
Minutes per session	0.025	0.007	3.68	(0.01, 0.04)
Trained therapist	0.338	0.237	1.43	(–0.18, 0.86)
Mean age	–0.007	0.007	–0.99	(–0.02, 0.01)
TRI study	–0.061	0.221	–0.28	(–0.55, 0.42)

NOTE: b_j = parameter estimate; $s_{\dot{w}}\, SE[b_j]$ = adjusted standard error of the parameter estimate; $t_{b_j} = b_j / (s_{\dot{w}}\, SE\,[b_j])$ (critical values = ± 2.20); estimate of residual heterogeneity: $\hat{\tau}^2_{ME} = 0$; test whether at least one moderator is significant: $Q_R = 28.91$ (critical value = 13.43); test for residual heterogeneity: $Q_E = 9.05$ (critical value = 19.68).

of results could occur by chance, but a more likely explanation is that the effectiveness of massage therapy depends on (or is moderated by) the duration of the treatment. So-called *forest plots,* such as the one shown in Figure 31.2, can be useful devices for concisely displaying the results from a meta-analysis and revealing interesting trends.

The influence of treatment duration is also apparent after plotting the effect size estimates against minutes of therapy provided, as shown in Figure 31.3. Circles represent effect size estimates from studies where a trained therapist provided the treatment, while squares represent effect size estimates from studies with a layperson. Moreover, larger points correspond to effect size estimates with smaller sampling variances. The lines indicate the estimated effect sizes as a function of minutes of therapy provided to a sample with an average age of 40 in a non-TRI study, once for a layperson and once for a trained therapist. The lines were plotted separately just for illustration purposes since the moderator distinguishing between a layperson and a trained therapist providing the treatment was not statistically significant.

Other Models as Special Cases of the Mixed Effects Model

As discussed earlier, the fixed effects with moderators, the random effects, and the fixed effects models are all special cases of the mixed effects model. Therefore, these models are applicable depending on the fit of the mixed effects model and the results from a moderator analysis.

Fixed Effects With Moderators Model

The mixed effects model reduces to the fixed effects with moderators model when $\tau^2_{ME} = 0$ (cf. Equations 5 and 7). Therefore, when the estimate of τ^2_{ME} is zero, this indicates that no residual heterogeneity is present, and the fixed effects with moderators model applies. This is exactly what happened in the illustrative example since $\hat{\tau}^2_{ME} = 0$. Therefore, the mixed effects model we fitted earlier actually corresponds to a fixed effects with moderators model. In general, then, to fit a fixed effects with moderators model, we simply need to apply all of the equations given earlier, except that w_i is always set equal to $1/\hat{v}_i$.

Random Effects Model

When none of the moderators included in the model influence the effect sizes (i.e., we conclude that $\beta_1 = \ldots = \beta_p = 0$), but heterogeneity is present (i.e., $\hat{\tau}^2_{ME} > 0$), then this suggests either that the heterogeneity in the effect sizes is entirely random (and could not be accounted for, no matter which set of moderators is included in the model) or that the heterogeneity is (at least in part) a result of moderators, but we lack the necessary information about the relevant moderators to account for it. In either case, the best we can usually do is to adopt the

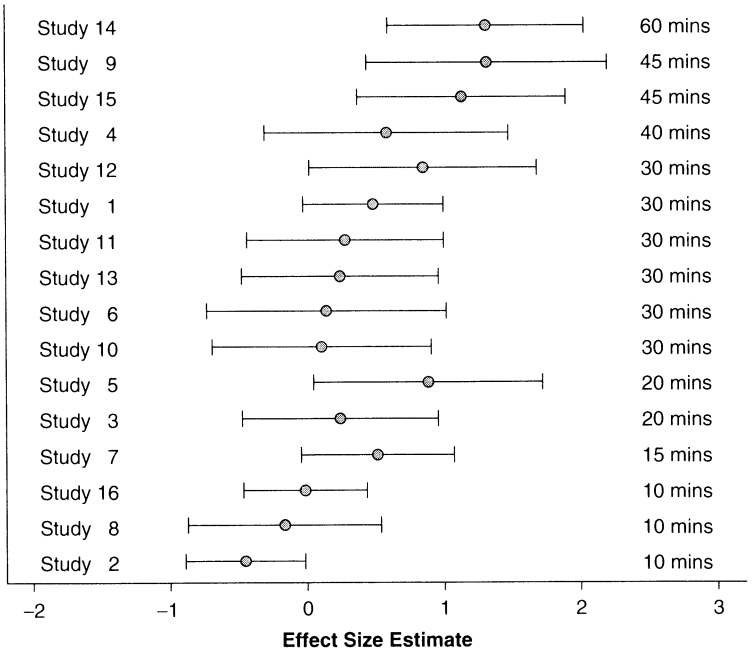

Figure 31.2 Individual effect size estimates (ordered by minutes of therapy provided per session) with corresponding 95% confidence intervals.

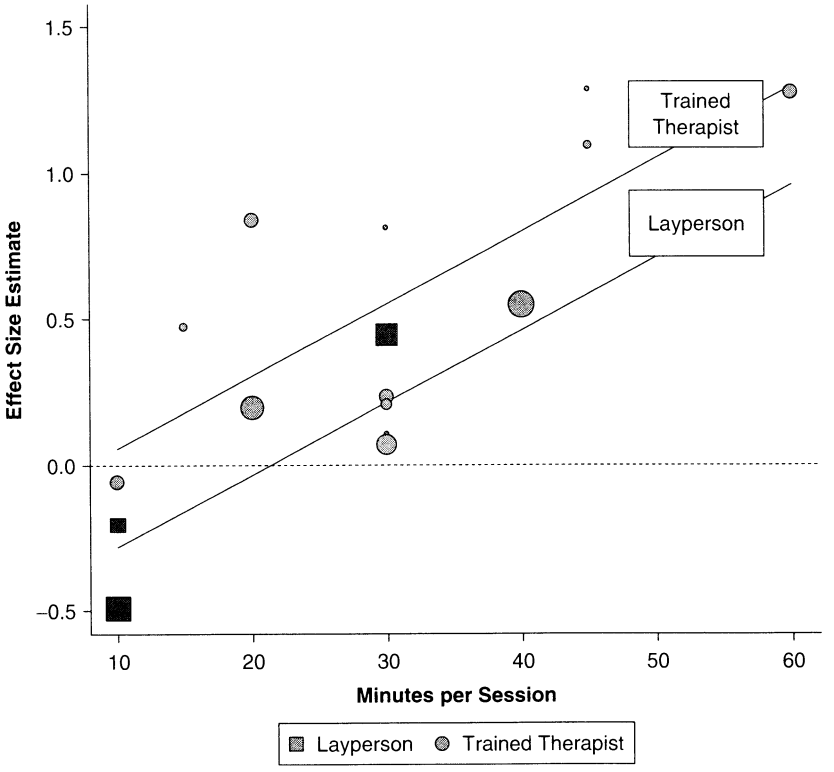

Figure 31.3 Effect size estimates and estimated effect sizes as a function of minutes of therapy provided to a sample with an average age of 40 in a non-TRI study, once for a layperson and once for a trained therapist.

random effects model (Equation 6) and treat the heterogeneity as purely random.

Fitting the random effects model requires that we estimate μ, the average effect size, and τ^2_{RE}, the amount of heterogeneity in the effect sizes. First, we estimate τ^2_{RE} with

$$\hat{\tau}^2_{RE} = \frac{Q - (k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}, \tag{16}$$

where $w_i = 1/\hat{v}_i$,

$$Q = \sum w_i (Y_i - \hat{\theta})^2, \tag{17}$$

and

$$\hat{\theta} = \frac{\sum w_i Y_i}{\sum w_i}. \tag{18}$$

After obtaining $\hat{\tau}^2_{RE}$ we can estimate μ with

$$\hat{\mu} = \frac{\sum w_i Y_i}{\sum w_i}, \tag{19}$$

where $w_i = 1/(\hat{\tau}^2_{RE} + \hat{v}_i)$. The standard error of $\hat{\mu}$ is approximately equal to

$$SE[\hat{\mu}] = \sqrt{\frac{1}{\sum w_i}}. \tag{20}$$

Finally, we can test if the average effect size differs significantly from zero (i.e., $H_0 : \mu = 0$) by comparing

$$t = \frac{\hat{\mu}}{s_{\hat{w}} SE[\hat{\mu}]} \tag{21}$$

against the critical values of a t distribution with $k-1$ degrees of freedom, where $s^2_{\hat{w}}$ is the larger of

$$s^2_{\hat{w}} = \frac{\sum w_i (Y_i - \hat{\mu})^2}{k-1} \tag{22}$$

and 1. Alternatively, a confidence interval for μ can be constructed with

$$\hat{\mu} \pm t_{(k-1;1-\alpha/2)} s_{\hat{w}} SE[\hat{\mu}]. \tag{23}$$

do not truncate to 1 (just leave (22) as is, even if it below 1).

Illustrative Example

Although the data strongly suggest that the heterogeneity in the effect sizes is not random, we now fit the random effects model to the data for illustration purposes. First, we calculate the value of $\hat{\theta}$ (Equation 18), which is equal to 0.280. We then apply Equation 17, which yields $Q = 37.959$. From this, we can obtain the estimate of τ^2_{RE} (Equation 16), which is equal to 0.175. This value is now an estimate of the total amount of heterogeneity in the effect sizes (as opposed to the estimate of the residual amount of heterogeneity we obtained earlier when fitting the mixed effects model). The average effect size μ, estimated with Equation 19, is equal to $\hat{\mu} = 0.379$ ($SE[\hat{\mu}] = 0.138$). Equation 22 then yields a value of 0.912 for $s^2_{\hat{w}}$, which is below 1 and therefore set to 1. The value of t (Equation 21) is equal to 2.75, and a 95% confidence interval for μ is given by (0.09, 0.67). This interval excludes zero, which indicates that, on average, massage therapy is effective for reducing state anxiety.

do not truncate to 1 (this is overly conservative)

Fixed Effects Model

Should we find that $\hat{\tau}^2_{RE} = 0$, this would provide evidence that the effect sizes are homogeneous. In other words, $\hat{\tau}^2_{RE} = 0$ suggests that neither moderators nor an additional source of random variability are influencing the effect sizes. Equation 18 then yields the estimate of θ, the homogeneous effect size for all studies. The standard error of the estimate, the statistic for testing whether $H_0 : \theta = 0$, and a confidence interval for θ can then be obtained with Equations 20 through 23, replacing $\hat{\mu}$ with $\hat{\theta}$ and setting $w_i = 1/\hat{v}_i$ in all of the equations. The application of the fixed effects model will not be illustrated with the example since the data clearly indicate that the effect sizes are not homogeneous.

Testing for the Presence of Residual Heterogeneity

A common practice in meta-analysis is to test whether the estimated amount of residual heterogeneity in a mixed effects model (i.e., $\hat{\tau}^2_{ME}$) is significantly greater than zero. To test the null hypothesis $H_0 : \tau^2_{ME} = 0$, we fit the fixed effects with moderators model (i.e., we set $w_i = 1/\hat{v}_i$ and use Equations 9 and 10 to obtain $\mathbf{b}$ and $\hat{\Sigma}$) and then calculate

$$Q_E = \sum w_i Y_i^2 - \mathbf{b}' \hat{\Sigma}^{-1} \mathbf{b}. \tag{24}$$

If Q_E exceeds the critical value of a chi-square random variable with $k - p - 1$ degrees of freedom, we conclude that there is additional heterogeneity in the effect sizes that is not accounted for by the moderators included in the model. This might indicate the presence of other moderators that we have missed, additional random heterogeneity, or both.

Illustrative Example

Since we found $\tau^2_{ME} = 0$ in the example given earlier, we have already fitted the fixed effects with moderators model to our data. Therefore, the Q_E statistic can be directly computed with $\mathbf{b}$ and $\hat{\Sigma}$ as given above and is equal to 9.05. Compared against 19.68, the critical value of a chi-square random variable with 16 – 4 – 1 = 11 degrees of freedom, we conclude that no residual heterogeneity is present. This is not a surprising finding since the estimate of residual heterogeneity was zero.

Confidence Interval for the Amount of Residual Heterogeneity

Instead of (or in addition to) testing whether the amount of residual heterogeneity is equal to zero, one can also report a confidence interval for τ^2_{ME}. The most accurate method to obtain such a confidence interval works as follows (Viechtbauer, 2007a). Let $Q_E(\tilde{\tau}^2_{ME})$ denote the value of Equation 24 when setting $w_i = 1/(\tilde{\tau}^2_{ME} + \hat{v}_i)$ (note that $\mathbf{b}$ and $\hat{\Sigma}$ also need to be recalculated to obtain this value). Moreover, let $\chi^2_{k-p-1;\,0.025}$ and $\chi^2_{k-p-1;\,0.975}$ denote the 2.5th and 97.5th percentiles of a chi-square distribution with $k-p-1$ degrees of freedom. Then the lower and upper bounds of a 95% confidence interval for τ^2_{ME} are given by those two $\tilde{\tau}^2_{ME}$ values, where $Q_E(\tilde{\tau}^2_{ME}) = \chi^2_{k-p-1;\,0.975}$ and $Q_E(\tilde{\tau}^2_{ME}) = \chi^2_{k-p-1;\,0.025}$. These values must be obtained iteratively. The simplest approach is to start with $\tilde{\tau}^2_{ME} = 0$ and to compute $Q_E(\tilde{\tau}^2_{ME})$ repeatedly for increasing $\tilde{\tau}^2_{ME}$ values until $Q_E(\tilde{\tau}^2_{ME})$ is equal to $\chi^2_{k-p-1;\,0.975}$ and then equal to $\chi^2_{k-p-1;\,0.025}$. If $Q_E(\tilde{\tau}^2_{ME})$ falls below $\chi^2_{k-p-1;\,0.975}$ for $\tilde{\tau}^2_{ME} = 0$, then the lower bound is set to zero. Moreover, if $Q_E(\tilde{\tau}^2_{ME})$ even falls below $\chi^2_{k-p-1;\,0.975}$ for $\tilde{\tau}^2_{ME} = 0$, then the lower and upper bounds are both below zero, and the confidence interval is equal to the null set.

Illustrative Example

With 11 degrees of freedom, $\chi^2_{11;\,0.975} = 21.92$ and $\chi^2_{11;\,0.025} = 3.82$. We have seen earlier that $Q_E = 9.05$, which is actually the value of $Q_E(\tilde{\tau}^2_{ME})$ for $\tilde{\tau}^2_{ME} = 0$. Therefore, the lower bound of a 95% confidence interval for $\tilde{\tau}^2_{ME}$ is 0. To obtain the upper bound, we increase $\tilde{\tau}^2_{ME}$ in small steps, each time recalculating $Q_E(\tilde{\tau}^2_{ME})$. For $\tilde{\tau}^2_{ME} = 0.185$, $Q_E(\tilde{\tau}^2_{ME}) = 3.82$. Therefore, a 95% confidence interval for τ^2_{ME} is given by (0, 0.185).

Testing for the Presence of Heterogeneity

We can also test whether the amount of heterogeneity in the random effects model is significantly greater than zero. If the amount of heterogeneity is zero, then this implies that the effect sizes are homogeneous. The null hypothesis is therefore given by H_0: $\theta_1 = \ldots = \theta_k$ (or, equivalently, $H_0 : \tau^2_{RE} = 0$). The statistic needed for this test is actually the one given in Equation 17. The null hypothesis is rejected when Q exceeds the critical value of a chi-square random variable with $k - 1$ degrees of freedom. In that case, we conclude that the effect sizes are heterogeneous, which might indicate the presence of moderators, random heterogeneity, or both.

Illustrative Example

We found earlier a value of $Q = 37.96$. The critical value of a chi-square random variable with 16 – 1 = 15 degrees of freedom is 25.00; therefore, we reject H_0 and conclude that the effect sizes are heterogeneous.

Confidence Interval for the Total Amount of Heterogeneity

Using the method described earlier, one can also obtain a confidence interval for τ^2_{RE} in the random effects model. Letting $\chi^2_{k-1;\,0.025}$ and $\chi^2_{k-1;\,0.975}$ denote the 2.5th and 97.5th percentiles of a chi-square distribution with $k-1$ degrees of freedom and

$$Q(\tilde{\tau}^2_{RE}) = \sum w_i (Y_i - \hat{\mu})^2, \qquad (25)$$

where $w_i = 1/(\tilde{\tau}^2_{RE} + \hat{v}_i)$ and $\hat{\mu}$ is calculated with Equation 19 after setting $w_i = 1/(\tilde{\tau}^2_{RE} + \hat{v}_i)$, we start with $\tilde{\tau}^2_{RE} = 0$ and iteratively increase $\tilde{\tau}^2_{RE}$ until we find those two $\tilde{\tau}^2_{RE}$ values, such that $Q(\tilde{\tau}^2_{RE}) = \chi^2_{k-1;\,0.975}$ and $Q(\tilde{\tau}^2_{RE}) = \chi^2_{k-1;\,0.025}$.

Illustrative Example

With 15 degrees of freedom, $\chi^2_{15;\,0.975} = 27.49$ and $\chi^2_{15;\,0.025} = 6.26$. We found that $Q = 37.96$, which is the value of $Q(\tilde{\tau}^2_{RE})$ for $\tilde{\tau}^2_{RE} = 0$. Increasing $\tilde{\tau}^2_{RE}$ slowly and recalculating $Q(\tilde{\tau}^2_{RE})$ each time reveals that $Q(\tilde{\tau}^2_{RE}) = 27.49$ when $\tilde{\tau}^2_{RE} = 0.034$ and $Q(\tilde{\tau}^2_{RE}) = 6.26$ when $\tilde{\tau}^2_{RE} = 0.524$. Therefore, a 95% confidence interval for τ^2_{RE} is given by (0.034, 0.524).

A Note About the Heterogeneity Tests

When we fail to reject the null hypothesis $H_0: \tau^2_{RE} = 0$ with the Q test, one should not automatically conclude that the effect sizes are truly homogeneous. The test lacks power to detect heterogeneity when k, the within-study sample sizes, or the amount of heterogeneity are small (Hunter & Schmidt, 2000; Sánchez-Meca & Marín-Martínez, 1997; Viechtbauer, 2007b). This, in turn, might lead researchers to adopt a fixed effects model too often, to miss the presence of moderators, or to attribute unwarranted precision to their results (National Research Council, 1992). Moreover, the Type I error rate of the Q test is only nominal when the within-study sample sizes are sufficiently large. In other words, when analyzing studies with small sample sizes, the test may be very inaccurate (Viechtbauer, 2007b). Therefore, a better approach would be to always adopt a random effects model. When the amount of heterogeneity is estimated to be zero (i.e., $\hat{\tau}^2_{RE} = 0$), then the random effects model simplifies to the fixed effects model anyway.

The same concerns apply to the Q_E test. In other words, one should not assume that residual heterogeneity is completely absent when we fail to reject $H_0: \tau^2_{ME} = 0$ with the Q_E test. Again, the better approach would be to always start with a mixed effects model, which will automatically reduce to a fixed effects with moderators model when $\hat{\tau}^2_{ME} = 0$ (as demonstrated with the example given earlier).

Quantifying the Amount of (Residual) Heterogeneity

Raw estimates of τ^2_{ME} and τ^2_{RE} are difficult to interpret. For example, in our example, we found that $\hat{\tau}^2_{RE} = 0.175$. Does this value indicate a small or large amount of heterogeneity among the effect sizes? To answer this question, it may be useful to express the amount of heterogeneity in terms of a value that is easier to interpret.

First note that the amount of variability among the effect size estimates can be decomposed into two parts: heterogeneity among the effect sizes (i.e., variability among the θ_i values) and sampling variability. The amount of sampling variability can be estimated by the $\hat{v}_i$ values, while the amount of heterogeneity among the effect sizes is estimated with $\hat{\tau}^2_{RE}$. Therefore, $\sum(\hat{\tau}^2_{RE} + \hat{v}_i)$ estimates the total amount of variability across the k effect size estimates. Consequently,

$$\hat{V}^2_T = \frac{k\hat{\tau}^2_{RE}}{\sum(\hat{\tau}^2_{RE} + \hat{v}_i)} \tag{26}$$

denotes the proportion of total variability in the effect size estimates that is due to heterogeneity (i.e., the proportion of variability in the effect size estimates that is not accounted for by sampling variability). An alternative method for estimating this quantity is discussed in Higgins, Thompson, Deeks, and Altman (2003).

Turning now to the amount of residual heterogeneity, first note that $\hat{\tau}^2_{ME}$ will tend to be smaller than $\hat{\tau}^2_{RE}$ if the moderator(s) included in the mixed effects model account for (at least some of) the heterogeneity among the effect sizes. Consequently,

$$\hat{V}^2_R = \frac{k\hat{\tau}^2_{ME}}{\sum(\hat{\tau}^2_{RE} + \hat{v}_i)} \tag{27}$$

denotes the proportion of total variability among the effect size estimates that is due to residual heterogeneity (i.e., not accounted for by sampling variability and the moderator[s] in the model). A value of $\hat{V}^2_R$ larger than 1 should be truncated to 1. Finally, we can also compute

$$\hat{R}^2 = \frac{\hat{\tau}^2_{RE} - \hat{\tau}^2_{ME}}{\hat{\tau}^2_{RE}} \tag{28}$$

as an estimate of the proportion of heterogeneity that is explained by the moderator(s) included in the model (Raudenbush, 1994). In rare cases, $\hat{R}^2$ may become negative, in which case it should be set to zero.

Illustrative Example

In the random effects model, we found that $\hat{\tau}^2_{RE} = 0.175$. Based on Equation 26, we then find that $\hat{V}^2_T = 0.56$, indicating that 56% of the total variability in the effect size estimates is due to heterogeneity (and therefore unaccounted for). On the other hand, $\hat{\tau}^2_{ME} = 0$ in the mixed effects model. It requires no further computation to see that $\hat{V}^2_R = 0$ and $\hat{R}^2 = 1$, indicating that the proportion of total variability unaccounted for is zero and that the proportion of heterogeneity accounted for by the moderators is 1.

Testing One Moderator at a Time Is Not Generally a Best Practice

While many authors choose to fit a separate model for each moderator variable of interest (instead of fitting a single mixed effects model that includes all moderators simultaneously), when moderators are correlated, this can cause drastic overestimation of moderator effects.[6] Taking the current example, Table 31.3 summarizes the effects when each moderator is analyzed individually as compared with simultaneously.

In terms of the statistical significance of the moderators, the conclusions from this approach are identical to the ones we obtained earlier (cf. Table 31.2). However, some of the parameter estimates have changed substantially. In particular, massage therapy from a fully trained therapist is now estimated to be 0.6 SMD units higher than when a layperson provides the treatment, almost double compared to what we found when fitting the mixed effects model with all moderators entered simultaneously (cf. Table 31.2).

The reason why this moderator now appears to have a stronger impact on the effect sizes can be explained based on the correlation among the moderators. Specifically, a layperson provided the therapy in Studies 1, 2, and 8, two of which (Studies 2 and 8) also happen to be studies where only 10 minutes of therapy were provided (see Figures 31.2 and 31.3). Since the effectiveness of therapy increases with treatment duration, the difference in the effectiveness between a layperson and a trained therapist is exacerbated when this moderator is examined by itself.

Note also that the sign of the TRI study moderator has changed. When this moderator is examined by itself, the data suggest that TRI studies yield SMDs that are 0.232 units larger than those in non-TRI studies. On the other hand, with all moderators entered simultaneously into the model, TRI studies are estimated to yield SMDs that are 0.061 units below those of non-TRI studies. In general, one may draw completely different conclusions from the analysis depending on the approach chosen.

Model Selection Strategy

To summarize and complete the recommendations given throughout this chapter, the following model selection strategy is suggested. First, an estimate of the total amount of heterogeneity (i.e., τ^2_{RE}) should be calculated when starting with the meta-analysis. This estimate can be supplemented with the results from the Q test

Table 31.3 Results From Fitting the Mixed Effects Model to the Data in Table 31.1 When Examining One Moderator at a Time

Moderator	b_j	$s_{\hat{w}}SE[b_j]$	t_{b_j}	*95% CI for* β_j	$\hat{\tau}^2_{ME}$	*95% CI for* τ^2	Q_E	$\hat{V}^2_R$	$\hat{R}^2$
Minutes per session	0.030	0.006	5.08	(0.02, 0.04)	0	(0, 0.133)	12.18	0	1
Trained therapist	0.600	0.293	2.04	(−0.03, 1.23)	0.121	(0.007, 0.422)	28.12	.39	.31
Mean age	−0.013	0.009	−1.35	(−0.03, 0.01)	0.153	(0.022, 0.516)	32.25	.49	.13
TRI study	0.232	0.271	0.86	(−0.35, 0.81)	0.166	(0.030, 0.568)	34.05	.53	.05

NOTE: b_j = parameter estimate; $s_{\hat{w}}SE[b_j]$ = adjusted standard error of parameter estimate; $t_{b_j} = b_j/s_{\hat{w}}SE[b_j]$ (critical values = ± 2.14); τ^2_{ME} = estimate of residual heterogeneity; Q_E = test for residual heterogeneity (critical value = 23.68); $\hat{V}^2_R$ = proportion of total variability in the effect size estimates due to residual heterogeneity; $\hat{R}^2$ = proportion of heterogeneity that is explained by the moderator.

and, for easier interpretation, given as a proportion relative to the total amount of variability (Equation 26). When τ^2_{RE} is estimated to be zero and the Q test is not significant, one has support for the hypothesis that the fixed effects model holds. However, when in doubt (such as when $\tau^2_{RE} > 0$, regardless of the results from the Q test), one should not adopt the fixed effects model.

When heterogeneity appears to be present, one can try to account for the heterogeneity by fitting a mixed effects model to the data. However, the number of potential moderator variables is usually quite large, especially when compared with the number of effect size estimates. This may lead to overfitting and increases the risk of finding significant moderators by chance alone. Prespecification of moderator variables based on expert knowledge and theoretical considerations is therefore a necessary prerequisite in most meta-analyses (Thompson & Higgins, 2002).

After fitting the mixed effects model, one can proceed with the moderator analysis as demonstrated earlier. A significant Q_R test (Equation 13) can be followed by individual moderator tests with the t_{b_j} statistic (Equation 14) and/or corresponding confidence intervals (Equation 15). Moreover, liberal use of plots and figures such as the ones shown in Figures 31.2 and 31.3 can greatly improve the interpretability of the results.

Due to missing data, it is often not possible to include multiple moderators in the mixed effects model simultaneously. Each study with missing data on any one of the moderator variables would have to be excluded from the model. In this case, one can fit the mixed effects model to each moderator variable separately. The Bonferroni correction may be used then to account for the fact that multiple hypothesis tests are being conducted. However, as demonstrated earlier, this approach is less than ideal, especially when the moderator variables are strongly correlated.[7]

In the unlikely event that none of the moderators appear to be related to the effect sizes, the best we can usually do is to treat the heterogeneity as completely unsystematic and adopt the random effects model. On the other hand, if a model is found that can account for all of the heterogeneity (i.e., the estimate of residual heterogeneity is zero), one automatically adopts the fixed effects with moderators model (as shown in the example given earlier). However, if the estimate of residual heterogeneity is greater than zero (regardless of the results from the Q_E test), the results from the mixed effects model should be reported.

Other Issues

There are many issues one may encounter while conducting a meta-analysis that are beyond the scope of the present chapter. For example, publication bias is a salient issue in meta-analysis and is discussed in Chapter 12 (see also Rothstein, Sutton, & Borenstein, 2005).

Dependent Effect Size Estimates

It was assumed throughout this chapter that the effect size estimates are independent. This assumption may be violated if multiple effect size estimates are obtained from the same sample of subjects. Methods for dealing with dependent effect size estimates can be found in Gleser and Olkin (1994), Kalaian and Raudenbush (1996), and Raudenbush, Becker, and Kalaian (1988).

Conclusions

Over the past three decades, meta-analysis has established itself as a viable approach for dealing with the ever increasing body of primary research. A quick search of the PsychINFO database revealed 17 citations involving the search term *meta-analysis* up to 1979, 918 citations between 1980 and 1989, 2,412 citations between 1990 and 1999, and already 2,418 citations between 2000 and 2005. The same search within the MEDLINE database revealed 1, 497, 5,851, and 9,622 citations involving that search term in the same intervals.

However, meta-analytic techniques currently employed in practice often lag behind recent methodological developments. Too much emphasis is still put on simple overall effects that do not take into account the heterogeneity typically present in the data (Lau et al., 1998). Models that allow for (residual) heterogeneity remain underused (Field, 2003; Hunter & Schmidt, 2000; National Research Council, 1992). Refined techniques for moderator analysis have been developed but appear to be largely unknown among practitioners.

Some of the current meta-analytic methods were introduced in the present chapter, with

particular emphasis on model fitting, model selection, and moderator analysis. While this chapter can only scratch the surface of the entire array of techniques available, it is hoped that it will help to make some of these techniques more accessible to the practitioner.

Notes

1. Several books have already been written that describe in detail the entire process from beginning to end (e.g., Cooper, 1998; Cooper & Hedges, 1994; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001; Rosenthal, 1991), and those planning to conduct a meta-analysis would be well advised to consult these sources.

2. Note that we could have considered other moderator variables as well, such as the instrument used in each study to measure anxiety levels (some studies used the State-Trait Anxiety Inventory, others used a visual analog scale, and yet others used measures constructed by the investigator), type of control group (in some studies, subjects in the control group received no treatment at all, while some form of alternative or placebo treatment was used in others), or the gender distribution of the subjects (the percentage of females in the studies ranged from 24% to 100%). However, for this example, we will concentrate on the four moderator variables given in Table 31.1.

3. A more complete discussion of these and other effect size measures is beyond the scope of the present chapter. The interested reader could consult, for example, Fleiss (1994), Lipsey and Wilson (2001), and Rosenthal (1994) for further information.

4. The mixed effects model is, in turn, a special case of a two-stage hierarchical linear model (e.g., Raudenbush & Bryk, 1985). Specifically, the Level 2 structure (for the effect sizes) is given by $\theta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + u_i$, while the Level 1 structure (for the effect size estimates) is given by Equation 3. For more information on hierarchical linear modeling, see Chapters 29–31.

5. By centering the moderator variables (in particular, the minutes of treatment and the mean age variables), one can make the intercept more interpretable. For example, subtracting 30 from the minutes of treatment variable and 40 from the mean age variable leaves all of the parameter estimates unchanged, except for the intercept, which is now equal to .214 and indicates the estimated effect for 30 minutes of treatment by a layperson to a sample with a mean age of 40 in a non-TRI study.

6. Editor's note: This issue is similar to running multiple simple regressions to assess the effect of multiple predictors or performing a single multiple regression. Few would argue that multiple simple regressions are superior to one multiple regression.

7. See Pigott (1994, 2001) for information on dealing with this issue. The more general issue of missing data is addressed in Chapter 15.

References

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25,* 12–37.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., et al. (1992). *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

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

Field, A. P. (2003). The problem in using fixed-effects models of meta-analysis on real world data. *Understanding Statistics, 2,* 105–124.

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

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York: Russell Sage Foundation.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

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

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

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

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research

knowledge. *International Journal of Selection and Assessment, 8,* 275–292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

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

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22,* 2693–2710.

Lau, J., Ioannidis, J. P. A., & Schmid, C. H. (1998). Summing up evidence: One answer is not always enough. *Lancet, 351,* 123–127.

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

Moyer, C. A., Rounds, J., & Hannum, J. W. (2004). A meta-analysis of massage therapy research. *Psychological Bulletin, 130,* 3–18.

National Research Council. (1992). *Combining information: Statistical issues and opportunities.* Washington, DC: National Academic Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379.

Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 163–175). New York: Russell Sage Foundation.

Pigott, T. D. (2001). Missing predictors in models of effect size. *Evaluation and the Health Professions, 24,* 277–307.

Raudenbush, S. W. (1994). Random effects models. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

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

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

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage.

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

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments.* Chichester, England: John Wiley.

Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. *Quality & Quantity, 31,* 385–399.

Sidik, K., & Jonkman, J. N. (2003). On constructing confidence intervals for a standardized mean difference in meta-analysis. *Communications in Statistics, Simulation and Computation, 32,* 1191–1203.

Sidik, K., & Jonkman, J. N. (2005). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics, 15,* 823–838.

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal, 309,* 1351–1355.

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21,* 1559-1573.

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine, 18,* 2693–2708.

Viechtbauer, W. (2007a). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26,* 37–52.

Viechtbauer, W. (2007b). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology, 32,* 39–60.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6,* 413–429.