CHAPTER 19

Subgroup Analyses

INTRODUCTION

To this point our focus has been primarily on estimating the mean effect, and in that context variation in effect sizes served primarily to qualify the mean effect. For example, we used the variation to assign weights for computing the mean effect, and to make projections about the distribution of true effect sizes about the mean effect.

Now, our focus shifts from the mean effect to the variation itself. In this chapter we show how meta-analysis can be used to compare the mean effect for different subgroups of studies (akin to analysis of variance in a primary study). In the next chapter we show how meta-analysis can be used to assess the relationship between study-level covariates and effect size (akin to multiple regression in primary studies).

Consider the following examples.

- We anticipate that a class of drugs reduces the risk of death in patients with cardiac arrhythmia, but we hypothesize that the magnitude of the effect depends on whether the condition is acute or chronic. We want to determine whether the drug is effective for each kind of patient, and also to determine whether the effect differs in the two.
- Our meta-analysis includes 10 studies that used proper randomization techniques and 10 that did not. Before computing a summary effect across all 20 studies we want to compute the effect for each group of 10, and determine if the effect size is related to the kind of randomization employed in the study.

Introduction to Meta-Analysis. Michael Borenstein, L. V. Hedges, J. P. T. Higgins and H. R. Rothstein
 ISBN: 978-0-470-05724-7

- We anticipate that forest management reduces the destruction of tree stands by insect pests, but we hypothesize that the magnitude of the effect depends on the diversity of trees in the stand. We want to determine whether forest management is effective in reducing destruction for both single species and mixed stands, and also to determine whether the effect differs in the two.
- We have data from ten studies that looked at the impact of tutoring on math scores of ninth-grade students. Five of the studies used one variant of the intervention while five used another variant. We anticipate that both variants are effective, and our primary goal in the analysis is to determine whether one is *more effective* than the other.

We shall pursue the last of these examples (the impact of tutoring on math) throughout this chapter. The effect size in this example is the standardized mean difference between groups (Hedges' g) but the same formulas would apply for any effect size index. As always, if we were working with odds ratios or risk ratios all values would be in log units, and if we were working with correlations all values would be in Fisher's z units.

Assume all the studies used the same design, with some students assigned to be tutored and others to a control condition. In some studies (here called A) students were tutored once a week while in the others (B) students were tutored twice a week. Our goal is to compare the impact of the two protocols to see if either intervention is more effective than the other.

Note. In this example we will be comparing the effect in one subgroup of studies versus the effect in a second subgroup of studies. The ideal scenario would be to have studies that directly compare the two variants of the intervention, since this would remove the potential for confounds and also reduce the error term. We assume that such studies are not available to us.

How this chapter is organized

We present three computational *models*. These are (a) fixed-effect, (b) random-effects using separate estimates of τ^2, and (c) random-effects using a pooled estimate of τ^2.

For each of the three models we present three *methods* for comparing the subgroups. These are (1) the Z-test, (2) a Q-test based on analysis of variance, and (3) a Q-test for heterogeneity.

The three statistical *models*, crossed with the three computational *methods*, yield a total of nine possible combinations. These are shown in Box 19.1, which serves as a roadmap for this chapter. Readers who want to get a sense of the issues quickly may find it easier to read the introduction and method 1 for each model, and return later to methods 2 and 3.

The dataset and all computations are available on the book's web site.

BOX 19.1 ROADMAP

	Introduction	Method 1	Method 2	Method 3
Model		*Z*-test	*Q*-test based on ANOVA	*Q*-test for heterogeneity
Fixed-effect	Page 151	Page 156	Page 157	Page 158
Random-effects with separate estimates of τ^2	Page 164	Page 167	Page 169	Page 170
Random-effects with pooled estimate of τ^2	Page 171	Page 176	Page 177	Page 178

FIXED-EFFECT MODEL WITHIN SUBGROUPS

A forest plot of the Tutoring studies is shown in Figure 19.1. The five *A* studies (at the top) have effect sizes (Hedges' *g*) in the approximate range of 0.10 to 0.50. The five *B* studies (below) have effect sizes in the approximate range of 0.45 to 0.75.

The combined effect for the *A* studies (represented by the first diamond) is 0.32 with a 95% confidence interval of plus/minus 0.11. The combined effect for the *B* studies (represented by the second diamond) is 0.61 with a 95% confidence interval of plus/minus 0.12. Our goal, then, is to compare these two effects.

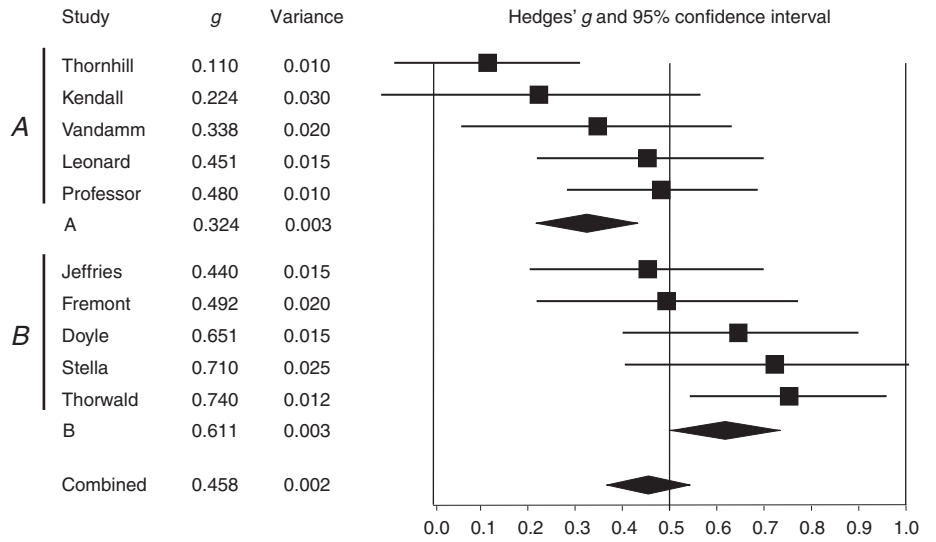

Figure 19.1 Fixed-effect model – studies and subgroup effects.

If we were working with a primary study (with Thornhill, Kendall, etc. being persons in treatment *A*, and Jeffries, Fremont, etc. being persons in treatment *B*), we would compute the mean and variance for each treatment, and our options for comparing these means would be clear. For example, we could perform a *t*-test to assess the difference between means relative to the standard error of the difference. Or, we could use analysis of variance to assess the variance among groups means relative to the variance within groups.

In meta-analysis we are working with subgroups of *studies* rather than groups of *subjects*, but will follow essentially the same approach, using a variant of the *t*-test or a variant of analysis of variance to compare the subgroup means. For this purpose we need to perform two tasks.

- Compute the mean effect and variance for each subgroup.
- Compare the mean effect across subgroups.

Computing the summary effects

In Table 19.1 the data for the *A* studies are displayed at the top, and data for the *B* studies are displayed toward the bottom.

To compute the summary effects we use the same formulas that we introduced for a single group (11.2) to (11.10). The summary effect for subgroup *A* is computed using values from the row marked *Sum A*. The summary effect for subgroup *B* is computed using values from the row marked *Sum B*. The summary effect for all studies is computed using values from the row marked *Sum*.

Table 19.1 Fixed effect model – computations.

	Study	Effect size Y	Variance Within V_Y	Variance Between T^2	Variance Total V	Weight W	Calculated quantities WY	WY^2	W^2
A	Thornhill	0.110	0.0100	0.0000	0.0100	100.000	11.000	1.210	10000.000
	Kendall	0.224	0.0300	0.0000	0.0300	33.333	7.467	1.673	1111.111
	Vandamm	0.338	0.0200	0.0000	0.0200	50.000	16.900	5.712	2500.000
	Leonard	0.451	0.0150	0.0000	0.0150	66.667	30.067	13.560	4444.444
	Professor	0.480	0.0100	0.0000	0.0100	100.000	48.000	23.040	10000.000
	Sum A					350.000	113.433	45.195	28055.556
B	Jefferies	0.440	0.0150	0.0000	0.0150	66.667	29.333	12.907	4444.444
	Fremont	0.492	0.0200	0.0000	0.0200	50.000	24.600	12.103	2500.000
	Doyle	0.651	0.0150	0.0000	0.0150	66.667	43.400	28.253	4444.444
	Stella	0.710	0.0250	0.0000	0.0250	40.000	28.400	20.164	1600.000
	Thorwald	0.740	0.0120	0.0000	0.0120	83.333	61.667	45.633	6944.444
	Sum B					306.667	187.400	119.061	19933.333
	Sum					656.667	300.833	164.255	47988.889

Computations (fixed effect) for the A studies

$$M_A = \frac{113.433}{350.000} = 0.3241,$$

$$V_{M_A} = \frac{1}{350.000} = 0.0029,$$

$$SE_{M_A} = \sqrt{0.0029} = 0.0535,$$

$$LL_{M_A} = 0.3241 - 1.96 \times 0.0535 = 0.2193,$$

$$UL_{M_A} = 0.3241 + 1.96 \times 0.0535 = 0.4289,$$

$$Z_A = \frac{0.3241}{0.0535} = 6.0633,$$

$$p(Z_A) < 0.0001,$$

$$Q_A = 45.195 - \left(\frac{113.433^2}{350.000}\right) = 8.4316 \tag{19.1}$$

$$p(Q = 8.4316,\ df = 4) = 0.0770,$$

$$C_A = 350.000 - \frac{28055.556}{350.000} = 269.8413,$$

$$T_A^2 = \frac{8.4316 - 4}{269.8413} = 0.0164,$$

and

$$I_A^2 = \left(\frac{8.4316 - 4}{8.4316}\right) \times 100 = 52.5594.$$

Computations (fixed effect) for the B studies

$$M_B = \frac{187.400}{306.667} = 0.6111,$$

$$V_{M_B} = \frac{1}{306.667} = 0.0033,$$

$$SE_{M_B} = \sqrt{0.0033} = 0.0571,$$

$$LL_{M_B} = 0.6111 - 1.96 \times 0.0571 = 0.4992,$$

$$UL_{M_B} = 0.6111 + 1.96 \times 0.0571 = 0.7230,$$

$$Z_B = \frac{0.6111}{0.0571} = 10.7013,$$

$$p(Z_B) < 0.0001,$$

$$Q_B = 119.011 - \left(\frac{187.400^2}{306.667}\right) = 4.5429, \tag{19.2}$$

$$p(Q = 4.5429, df = 4) = 0.3375,$$

$$C_B = 306.667 - \frac{19933.333}{306.667} = 241.667,$$

$$T_B^2 = \frac{4.5429 - 4}{241.667} = 0.0022,$$

and

$$I_B^2 = \left(\frac{4.5429 - 4}{4.5429}\right) \times 100 = 11.9506.$$

Computations (fixed effect) for all ten studies

$$M = \frac{300.833}{656.667} = 0.4581, \tag{19.3}$$

$$V_M = \frac{1}{656.667} = 0.0015, \tag{19.4}$$

$$SE_M = \sqrt{0.0015} = 0.0390,$$

$$LL_M = 0.4581 - 1.96 \times 0.0390 = 0.3816,$$

$$UL_M = 0.4581 + 1.96 \times 0.0390 = 0.5346,$$

$$Z = \frac{0.4581}{0.0390} = 11.7396,$$

$$p(Z) < 0.0001,$$

$$Q = 164.255 - \left(\frac{300.833^2}{656.667}\right) = 26.4371, \tag{19.5}$$

$$p(Q = 26.4371, df = 9) = 0.0017,$$

$$C = 656.667 - \frac{47988.889}{656.667} = 583.5871,$$

$$T^2 = \frac{26.4371 - 9}{538.5871} = 0.0299,$$

and

$$I^2 = \left(\frac{26.4371 - 9}{26.4371}\right) \times 100 = 65.96.$$

The statistics computed above are summarized in Table 19.2.

Table 19.2 Fixed-effect model – summary statistics.

	A	*B*	Combined
Y	0.3241	0.6111	0.4581
V	0.0029	0.0033	0.0015
SE_Y	0.0535	0.0571	0.0390
LL_Y	0.2193	0.4992	0.3816
UL_Y	0.4289	0.7230	0.5346
Z	6.0633	10.7013	11.7396
p2	0.0000	0.0000	0.0000
Q	8.4316	4.5429	26.4371
df	4.0000	4.0000	9.0000
p-value	0.0770	0.3375	0.0017
Numerator	4.4316	0.5429	17.4371
C	269.8413	241.6667	583.5871
T^2	0.0164	0.0022	0.0299
I^2	52.5594	11.9506	65.9569

Comparing the effects

If we return to Figure 19.1 and excerpt the diamonds for the two subgroups we get Figure 19.2. The mean effect size for subgroups *A* and *B* are 0.324 and 0.611, with variances of 0.003 and 0.003.

Figure 19.2 Fixed-effect – subgroup effects.

Our goal is to compare these two mean effects, and we describe three ways that we can proceed. These approaches are algebraically equivalent, and (it follows) yield the same *p*-value. Our goal in presenting three approaches is to provide insight into the process.

Comparing *A* versus *B*: a *Z*-test (Method 1)

Since there are only two subgroups here, we can work directly with the mean difference in effect sizes. In a primary study, if we wanted to compare the means in two groups we would perform a *t*-test. In meta-analysis the mean and variance are based on *studies* rather than *subjects* but the logic of the test is the same.

Concretely, let θ_A and θ_B be the true effects underlying groups A and B, let M_A and M_B be the estimated effects, and let V_{M_A} and V_{M_B} be their variances. If we use *Diff* to refer to the difference between the two effects, and elect to subtract the mean of A from the mean of B,

$$Diff = M_B - M_A,$$

the test statistic to compare the two effects is

$$Z_{Diff} = \frac{Diff}{SE_{Diff}}, \tag{19.6}$$

where

$$SE_{Diff} = \sqrt{V_{M_A} + V_{M_B}}. \tag{19.7}$$

Under the null hypothesis that the true effect size θ is the same for both groups,

$$H_0 : \theta_A = \theta_B, \tag{19.8}$$

Z_{Diff} would follow the normal distribution. For a two-tailed test, the *p*-value is given by

$$p = 2[1 - (\Phi(|Z|))], \tag{19.9}$$

where $\Phi(Z)$ is the standard normal cumulative distribution.

In the running example,

$$Diff = 0.6111 - 0.3241 = 0.2870,$$

$$SE_{Diff} = \sqrt{0.0029 \ + 0.0033} = 0.0782,$$

$$Z_{Diff} = \frac{0.2870}{0.0782} = 3.6691,$$

and

$$p = 2[1 - (\Phi(|3.6691|))] = 0.0002.$$

The two-tailed p-value corresponding to $Z_{Diff} = 3.6691$ is 0.0002. This tells us that the treatment effect is probably not the same for the A studies as for the B studies. In Excel, the function to compute a 2-tailed p-value for Z is =(1-(NORMSDIST (ABS(Z)))*2. Here, =(1-(NORMSDIST(ABS(3.6691))))*2 will return the value 0.0002.

Comparing *A* with *B*: a *Q*-test based on analysis of variance (Method 2)

In a primary study, the t-test can be used to compare the means in *two* groups, but to compare means in more than two groups we use analysis of variance. Concretely, we partition the total variance (of all subjects about the grand mean) into variance within groups (of subjects about the means of their respective groups) and variance between groups (of group means about the grand mean). We then test these various components of variance for statistical significance, with the last (variance between groups) addressing the hypothesis that effect size differs as function of group membership.

In meta-analysis the means are based on *studies* rather than *subjects* but the logic of the test is the same. Specifically, we compute the following quantities (where *SS* is the sum of squared deviations).

- Q_A, the weighted *SS* of all A studies about the mean of A.
- Q_B, the weighted *SS* of all B studies about the mean of B.
- $Q_{within,}$ the sum of Q_A and Q_B.
- $Q_{bet,}$ the weighted *SS* of the subgroup means about the grand mean.
- Q, the weighted *SS* of all effects about the grand mean.

We may write $Q_{within} = Q_A + Q_B$, to represent the sum of within-group weighted *SS*, or more generally, for p subgroups,

$$Q_{within} = \sum_{j=1}^{p} Q_j. \tag{19.10}$$

In the running example

$$Q_{within} = 8.4316 + 4.5429 = 12.9745. \tag{19.11}$$

The weighted *SS* are additive, such that $Q = Q_{within} + Q_{bet}$. Therefore, Q_{bet} can be computed as

$$Q_{bet} = Q - Q_{within}. \tag{19.12}$$

Under the null hypothesis that the effect size θ is the same for all groups, 1 to p, Q_{bet} would be distributed as chi-squared with degrees of freedom equal to $p - 1$.

In the running example,

$$Q_{bet} = 26.4371 - 12.9745 = 13.4626. \tag{19.13}$$

Table 19.3 Fixed-effect model – ANOVA table.

	Q	*df*	*p*	Formula
A	8.4316	4	0.0770	19.1
B	4.5429	4	0.3375	19.2
Within	12.9745	8	0.1127	19.11
Between	13.4626	1	0.0002	19.13
Total	26.4371	9	0.0017	19.5

Each Q statistic is evaluated with respect to the corresponding degrees of freedom. In the running example (Table 19.3),

- The 'Total' line tells us that for the full group of ten studies the variance is statistically significant ($Q = 26.4371$, $df = 9$, $p = 0.0017$).
- The 'Within' line tells us that the variance within groups (averaged across groups) is not statistically significant ($Q_{within} = 12.9745$, $df = 8$, $p = 0.1127$).
- The 'Between' line tells us that the difference between groups (the combined effect for A versus B) is statistically significant ($Q_{bet} = 13.4626$, $df = 1$, $p = 0.0002$), which means that the effect size is related to the frequency of tutoring.
- At a finer level of detail, neither the variance within subgroup A ($Q_A = 8.4316$, $df = 4$, $p = 0.0770$) nor within subgroup B ($Q_B = 4.5429$, $df = 4$, $p = 0.3375$) is statistically significant.

As always, the absence of statistical significance (here, within subgroups) means only that we cannot rule out the hypothesis that the studies share a common effect size, and it does not mean that this hypothesis has been proven.

In Excel, the function to compute a p-value for Q is =CHIDIST(Q,df). For the test of A versus B, =CHIDIST(13.4626,1) returns 0.0002.

Comparing *A* versus *B*: a *Q*-test for heterogeneity (Method 3)

The test we just described can be derived in a different way. We can think of the effect sizes for subgroups A and B as single studies (if we extract the two subgroup lines and the total line from Figure 19.1 and replace the diamonds with squares, to represent these as if they were studies, we get Figure 19.3). Then, we can test these 'studies' for heterogeneity, using precisely the same formulas that we introduced earlier (Chapter 16) to test the dispersion of single studies about the summary effect.

Concretely, we start with two 'studies' with effect sizes of 0.324 and 0.611, and variance of 0.003 and 0.003. Then, we apply the usual meta-analysis methods to compute Q (see Table 19.4).

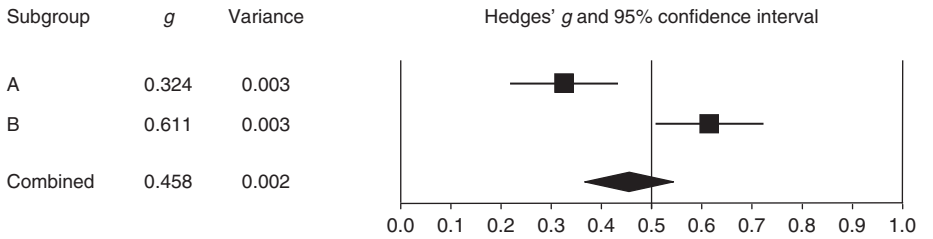

Figure 19.3 Fixed-effect model – treating subgroups as studies.

Table 19.4 Fixed-effect model – subgroups as studies.

Study	Effect size Y	Variance Within V_Y	Variance Between T^2	Variance Total V	Weight W	Calculated quantities		
						WY	WY^2	W^2
A	0.3241	0.0029	0.0000	0.0029	350.000	113.433	36.763	122500.000
B	0.6111	0.0033	0.0000	0.0033	306.667	187.400	114.518	94044.444
					656.667	300.833	151.281	216544.444

In this example,

$$M = \frac{300.833}{656.667} = 0.4581, \tag{19.14}$$

$$V_M = \frac{1}{656.667} = 0.0015, \tag{19.15}$$

$$Q = 151.281 - \left(\frac{300.833^2}{656.667}\right) = 13.4626,$$

$$df = 2 - 1 = 1,$$

and

$$p(Q = 13.4626, df = 1) = 0.0002,$$

where Q represents the weighted sum of squares for studies A and B about the grand mean. For $Q = 13.4626$ and $df = 1$, the p-value is 0.0002.

In Excel, the function to compute a p-value for Q is =CHIDIST(Q,df), and =CHIDIST(13.4626,1) returns 0.0002.

Summary

We presented three methods for comparing the effect size across subgroups. One method was to use a *Z*-test to compare the two effect sizes directly. Another was to use a *Q*-test to partition the variance, and test the between-subgroups portion of the variance. A third was to use a *Q*-test to assess the dispersion of the summary effects about the combined effect. All the methods assess the difference in subgroup effects relative to the precision of the difference (or the variance across subgroups effects relative to the variance within subgroups).

As noted earlier, the methods are mathematically equivalent. The two methods that report *Q*, report the same value for *Q* (13.4626). When there is one degree of freedom (so that we can use either a *Z*-test or a *Q*-test) *Z* is equal to the square root of *Q*. In our example, the method that reports *Z*, reports a value of $Z = 3.6691$, which is equal to the square root of *Q*. All three methods yield a *p*-value of 0.0002.

Quantify the magnitude of the difference

The *Z*-test and the *Q*-tests address the question of *statistical*, rather than *clinical* significance. In addition to reporting the test of significance, one should generally report an estimate of the effect size, which in this context is the difference in mean effect between the two subgroups. For subgroups *A* and *B*, if we elect to subtract the mean of *A* from the mean of *B*, the difference is

$$Diff = M_B - M_A. \tag{19.16}$$

The 95% confidence interval is estimated by

$$LL_{Diff} = Diff - 1.96 \times SE_{Diff} \tag{19.17}$$

and

$$UL_{Diff} = Diff + 1.96 \times SE_{Diff}, \tag{19.18}$$

where the standard error was defined in (19.7). If we had more than two subgroups, we could repeat this procedure for all pairs of subgroups. In the running example the difference in effects (which we have defined as *B* minus *A*) and its 95% confidence interval are estimated as

$$Diff = 0.6111 - 0.3241 = 0.2870,$$

$$SE_{Diff} = \sqrt{0.0029 + 0.0033} = 0.0782,$$

$$LL_{Diff} = 0.2870 - 1.96 \times 0.0782 = 0.1337,$$

and

$$UL_{Diff} = 0.2870 + 1.96 \times 0.0782 = 0.4403.$$

In words, the true difference between the effect in the subgroup *A* studies, as opposed to the subgroup *B* studies, probably falls in the range of 0.13 to 0.44.

COMPUTATIONAL MODELS

In Part 3 of this volume we discussed the difference between a fixed-effect model and a random-effects model. Under the fixed-effect model we assume that the true effect is the same in all studies. By contrast, under the random-effects model we allow that the true effect may vary from one study to the next. This difference has implications for the way that weights are assigned to the studies, which affects both the summary effect and its standard error.

When we introduced these two models we were working with a single set of studies. Now, we are working with more than one subgroup of studies (in the running example, *A* and *B*) but the same issues apply. Under the fixed-effect model we assume that all studies in subgroup *A* share a common effect size and that all studies in subgroup *B* share a common effect size. By contrast, under the random-effects model we allow that there may be some true variation of effects within the *A* studies and within the *B* studies.

When we initially discussed the fixed-effect model we used the example of a pharmaceutical company that enrolled 1000 patients for a clinical trial and divided them among ten cohorts of 100 patients each (page 83). These ten cohorts were known to be identical in all important respects, and so it was reasonable to assume that the true effect would be the same for all ten studies. When we presented this example we noted that the conditions described (of all the studies being performed by the same researchers using the same population and methods) are rare in systematic reviews, and that in most cases the random-effects model will be more plausible than the fixed-effect.

We can expand the pharmaceutical example to apply to subgroups if we assume that five of the studies will compare *Drug A* versus placebo, and the other five will compare *Drug B* versus placebo. Within the five *Drug A* studies and within the five *Drug B* studies there should be a single true effect size, and so in this case it would be correct to use the fixed-effect model within subgroups. However, the same caveat applies here, in that this kind of systematic review, where all studies are performed by the same researchers using the same population and methods, is very rare. In the vast majority of systematic reviews these conditions will not hold, and a random-effects analysis would be a better fit for the data.

For example, in the tutoring analysis it seems plausible that the distinction between the two interventions (one hour versus two hours a week) captures some, *but not all*, of the true variation among effects. Within either subgroup of studies (*A* or *B*) there are probably differences from study to study in the motivation of the students, or the dedication of the teachers, the details of the protocol, or other factors, such that the true effect differs from study to study. If these differences do

exist, and can have an impact on the effect size, then the random-effects model is a better match than the fixed-effect.

When we use the random-effects model, the impact on the summary effect within subgroups will be the same as it had been when we were working with a single population. The weights assigned to each study will be more moderate than they had been under the fixed-effect model (large studies will lose impact while small studies gain impact). And, the variance of the combined effect will increase.

T^2 should be computed within subgroups

To apply the random-effects model we need to estimate the value of τ^2, the variance of true effect sizes across studies. Since τ^2 is defined as the true variance in effect size among a set of studies, its value will differ depending on how we define the set.

If we were to define the set as all studies irrespective of which subgroup they belong to, with τ^2 based on the dispersion of all studies from the grand mean, τ^2 would tend to be relatively large. By contrast, if we define the set as all studies *within* a subgroup, with τ^2 based on the dispersion of the *A* studies from the mean of *A* and of the *B* studies from the mean of *B*, τ^2 would tend to be relatively small (especially if the *A* studies and the *B* studies do represent distinct clusters, as we have hypothesized).

Since our goal is to estimate the mean and sampling distribution of subgroup *A*, and to do the same for subgroup *B*, it is clearly the variance *within* subgroups that is relevant in the present context. Put simply, if some of the variance in effect sizes can be explained by the type of intervention, then this variance is not a factor in the sampling distribution of studies within a subgroup (where only one intervention was used). Therefore, we always estimate τ^2 within subgroups.

To pool or not to pool

When we estimate τ^2 within subgroups of studies, the estimate is likely to differ from one subgroup to the next. In the running example, the estimate of τ^2 in subgroup *A* was 0.016, while in subgroup *B* it was 0.002. We have the option to pool the within-group estimates of τ^2 and apply this common estimate to all studies. Alternatively, we can apply each subgroup's estimate of τ^2 to the studies in that subgroup.

Note. As a shorthand we refer to pooling the estimates of τ^2. In fact, though, what we actually pool are Q, df, and C, and then estimate τ^2 from these pooled values (see (19.38)).

The decision to pool (or not) depends on the following. If we assume that the true study-to-study dispersion is the same within all subgroups, then observed differences in T^2 must be due to sampling variation alone. In this case, we should pool the information to yield a common estimate, and then apply this estimate to all subgroups. This seems like a plausible expectation in the running example, where the study-to-study variation in effect size is likely to be similar for subgroups *A* and *B*.

On the other hand, if we anticipate that the true between-studies dispersion may actually differ from one subgroup to the next, then we would estimate τ^2 within subgroups and use a separate estimate of τ^2 for each subgroup. For example, suppose that we are assessing an intervention to reduce recidivism among juvenile delinquents, and comparing the effect in subgroups of studies where the delinquents did, or did not, have a history of violence. We might expect to see a wider range of effect sizes in one subgroup than the other.

There is one additional caveat to consider. If we do anticipate that τ^2 will vary from one subgroup to the next, so that the correct approach is to use separate estimates of τ^2, we still need to be sure that there are enough studies within each subgroup to yield an acceptably accurate estimate of τ^2. Generally, if there are only a few studies within subgroups (say, five or fewer), then the estimates of τ^2 within subgroups are likely to be imprecise. In this case, it makes more sense to use a pooled estimate, since the increased accuracy that we get by pooling more studies is likely to exceed any real differences between groups in the true value of τ^2.

Summary

The logic outlined above is encapsulated in the flowchart shown in Figure 19.4. If the studies within each subgroup share a common effect size, then we use the fixed-effect model to assign weights to each study (and τ^2 is zero). Otherwise, we use the random-effects model.

Under random effects we always estimate τ^2within subgroups. If we believe that the true value of τ^2 is the same for all subgroups, then the correct procedure is to pool the estimates obtained within subgroups. If we believe that the true value of τ^2 varies from one subgroup to the next, the correct procedure is to use a separate

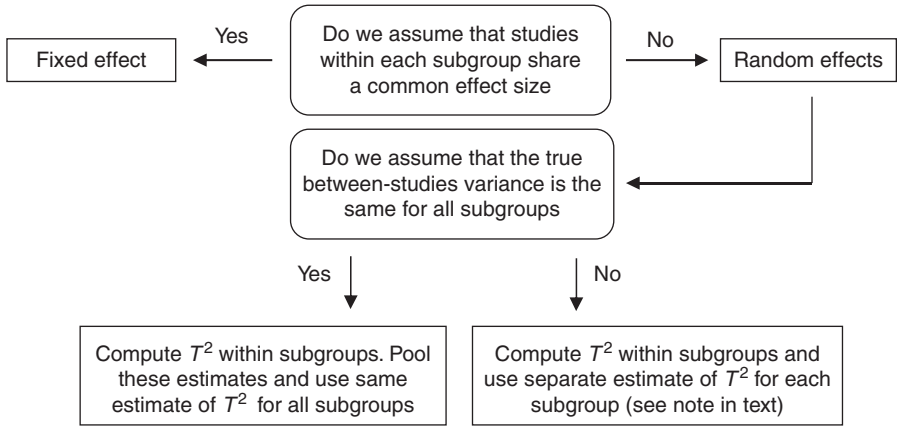

Figure 19.4 Flowchart for selecting a computational model.

estimate for each subgroup. However, if we have only a few studies within subgroups these estimates may be imprecise and therefore it may be preferable to pool the estimates.

RANDOM EFFECTS WITH SEPARATE ESTIMATES OF τ^2

Here, we proceed through the same set of computations as we did for the fixed-effect model, but this time using random-effects weights, with a separate estimate of τ^2 for each subgroup.

Computing the effects

Figure 19.5 is a forest plot of the studies in subgroups *A* and *B*. The studies are identical to those in the fixed-effect forest plot (Figure 19.2) but the summary effects, represented by the diamonds, are now based on random-effects weights. The mean effect size for subgroups *A* and *B* are 0.325 and 0.610, with variances of 0.006 and 0.004.

Computations are based on the values in Table 19.5. These values are similar to those in Table 19.1, except that the variance for each study now includes the within-study variance and the between-study variance. We did not assume a common value of τ^2 and therefore used a separate estimate of τ^2 for each subgroup. In Figure 19.5 this is indicated by the symbols at the right, where we have one value for T^2_A and another for T^2_B. In Table 19.5, the column labeled T^2 shows 0.0164 for the *A* studies and 0.0022 for the *B* studies.

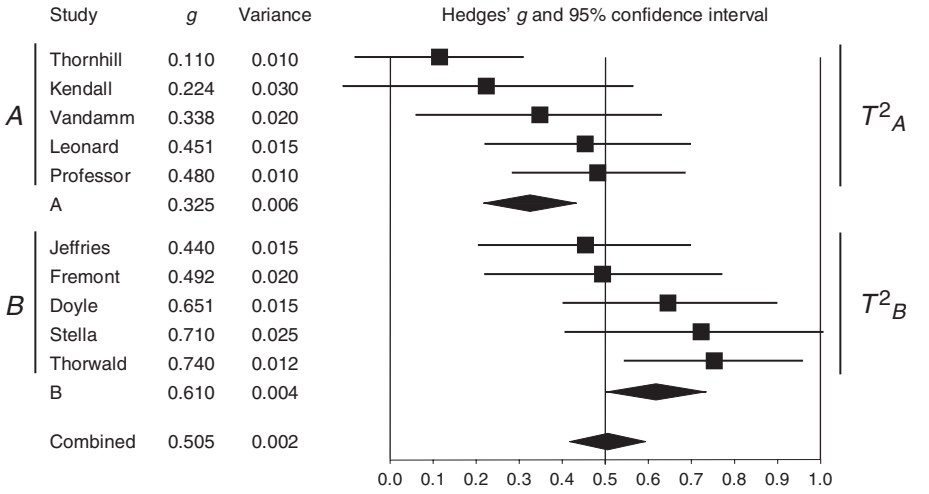

Figure 19.5 Random-effects model (separate estimates of τ^2) – studies and subgroup effects.

Table 19.5 Random-effects model (separate estimates of τ^2) – computations.

	Study	Effect size Y	Variance Within V_Y	Variance Between T^2	Variance Total V	Weight W	Calculated quantities WY	WY^2	W^2
A	Thornhill	0.110	0.0100	0.0164	0.0264	37.846	4.163	0.458	1432.308
	Kendall	0.224	0.0300	0.0164	0.0464	21.541	4.825	1.081	464.017
	Vandamm	0.338	0.0200	0.0164	0.0364	27.455	9.280	3.137	753.788
	Leonard	0.451	0.0150	0.0164	0.0314	31.824	14.353	6.473	1012.757
	Professor	0.480	0.0100	0.0164	0.0264	37.846	18.166	8.720	1432.308
	Sum A					156.512	50.787	19.868	5095.179
B	Jefferies	0.440	0.0150	0.0022	0.0172	57.983	25.512	11.225	3362.002
	Fremont	0.492	0.0200	0.0022	0.0222	44.951	22.116	10.881	2020.582
	Doyle	0.651	0.0150	0.0022	0.0172	57.983	37.747	24.573	3362.002
	Stella	0.710	0.0250	0.0022	0.0272	36.702	26.058	18.501	1347.034
	Thorwald	0.740	0.0120	0.0022	0.0142	70.193	51.943	38.438	4927.012
	Sum B					267.811	163.376	103.619	15018.633
	Sum					424.323	214.163	123.487	20113.812

Computations (random effects, separate estimates of τ^2) for the A studies

$$M_A^* = \frac{50.787}{156.512} = 0.3245,$$

$$V_{M_A^*} = \frac{1}{156.512} = 0.0064,$$

$$SE_{M_A^*} = \sqrt{0.0064} = 0.0799,$$

$$LL_{M_A^*} = 0.3245 - 1.96 \times 0.0799 = 0.1678,$$

$$UL_{M_A^*} = 0.3245 + 1.96 \times 0.0799 = 0.4812,$$

$$Z_A^* = \frac{0.3245}{0.0799} = 4.0595,$$

$$p(Z_A^*) < 0.0001,$$

and

$$Q_A^* = 19.868 - \left(\frac{50.787^2}{156.512}\right) = 3.3882. \tag{19.19}$$

Note. The Q^* statistic computed here, using random-effects weights, is used *only* for the analysis of variance, to partition Q^* into its various components. Therefore, we do not show a p-value for Q^*. Rather, the Q statistic computed using fixed-effect weights (Table 19.2) is the one that reflects the between-studies dispersion, provides a test of homogeneity for the studies within subgroup A, and is used to estimate T^2.

Computations (random effects, separate estimates of τ^2) for the B studies

$$M_B^* = \frac{163.376}{267.811} = 0.6100,$$

$$V_{M_B^*} = \frac{1}{267.811} = 0.0037,$$

$$SE_{M_B^*} = \sqrt{0.0037} = 0.0611,$$

$$LL_{M_B^*} = 0.6100 - 1.96 \times 0.0611 = 0.4903,$$

$$UL_{M_B^*} = 0.6100 + 1.96 \times 0.0611 = 0.7298,$$

$$Z_B^* = \frac{0.6100}{0.0611} = 9.9833,$$

$$p(Z_B^*) < 0.0001,$$

and

$$Q_B^* = 103.619 - \left(\frac{163.376^2}{267.811}\right) = 3.9523. \tag{19.20}$$

Computations (random effects, separate estimates of τ^2) for all ten studies

The statistics here are computed using the same value of T^2 as was used within groups (in this case, *not* pooled).

$$M^* = \frac{214.163}{424.323} = 0.5047, \tag{19.21}$$

$$V_{M^*} = \frac{1}{424.323} = 0.0024, \tag{19.22}$$

$$SE_{M^*} = \sqrt{0.0024} = 0.0485,$$

$$LL_{M^*} = 0.5047 - 1.96 \times 0.0485 = 0.4096,$$

$$UL_{M^*} = 0.5047 + 1.96 \times 0.0485 = 0.5999,$$

$$Z^* = \frac{0.5047}{0.0485} = 10.3967,$$

$$p(Z^*) < 0.0001,$$

and

$$Q^* = 123.487 - \left(\frac{214.163^2}{424.323}\right) = 15.3952. \tag{19.23}$$

Statistics (random-effects) are summarized in Table 19.6.

Table 19.6 Random-effects model (separate estimates of τ^2) – summary statistics.

	A	B	Combined
Y	0.3245	0.6100	0.5047
V	0.0064	0.0037	0.0024
SE_Y	0.0799	0.0611	0.0485
LL_Y	0.1678	0.4903	0.4096
UL_Y	0.4812	0.7298	0.5999
Z	4.0595	9.9833	10.3967
p2	0.0000	0.0000	0.0000
Q	3.3882	3.9523	15.3952

Comparing the effects

If we return to Figure 19.5 and excerpt the diamonds for the two subgroups we get Figure 19.6.

The mean effect size for subgroups *A* and *B* are 0.325 and 0.610, with variances of 0.006 and 0.004.

Our goal is to compare these two mean effects, and there are several ways that we can proceed. These approaches are algebraically equivalent, and (it follows) yield the same p-value. Our goal in presenting several approaches is to provide insight into the process.

Comparing *A* versus *B*: a *Z*-test (Method 1)

We can use a simple *Z*-test to compare the mean effect for subgroups *A* versus *B*. The formulas are identical to those used earlier, but we change two symbols to

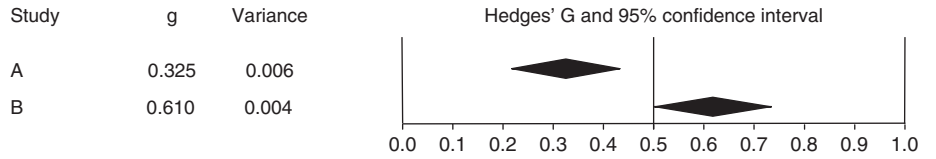

Figure 19.6 Random-effects model (separate estimates of τ^2) – subgroup effects.

reflect the random-effects model. First, we use a (*) to indicate that the statistics are based on random-effects weights rather than fixed-effect weights. Second, the null hypothesis is framed as $\mu_A = \mu_B$, reflecting the fact that these are mean values, rather than $\theta_A = \theta_B$, which we used to refer to common values when we were working with the fixed-effect model.

Let μ_A and μ_B be the true mean effects underlying subgroups A and B, let M_A^* and M_B^* be the estimated effects, and let $V_{M_A^*}$ and $V_{M_B^*}$ be their variances. If we use $Diff^*$ to refer to the difference between the two effects and elect to subtract the mean of A from the mean of B,

$$Diff^* = M_B^* - M_A^*. \tag{19.24}$$

The test statistic to compare the two effects is

$$Z_{Diff}^* = \frac{Diff^*}{SE_{Diff^*}}, \tag{19.25}$$

where

$$SE_{Diff^*} = \sqrt{V_{M_A^*} + V_{M_B^*}}. \tag{19.26}$$

Under the null hypothesis that the true mean effect size μ is the same for both groups,

$$H_0^* : \mu_A^* = \mu_B^*, \tag{19.27}$$

Z_{Diff}^* would follow the normal distribution. For a two-tailed test the p-value is given by

$$p^* = 2\left[1 - \left(\Phi\left(|Z_{Diff}^*|\right)\right)\right], \tag{19.28}$$

where $\Phi(Z)$ is the standard normal cumulative distribution.

In the running example

$$Diff^* = 0.6100 - 0.3245 = 0.2856,$$

$$SE_{Diff^*} = \sqrt{0.0064 + 0.0037} = 0.1006,$$

and

$$Z_{Diff^*} = \frac{0.2856}{0.1006} = 2.8381.$$

The two-tailed p-value corresponding to $Z_{Diff}^* = 2.8381$ is 0.0045. This tells us that the mean treatment effect is probably not the same for the A studies as for the B studies. In Excel, the function to compute a 2-tailed p-value for Z is =(1-(NORMSDIST(ABS(Z))))*2. Here, =(1-(NORMSDIST(ABS(2.8381))))*2 will return the value 0.0045.

Comparing *A* with *B*: a *Q*-test based on analysis of variance (Method 2)

We use the same formulas as we did for method 2 under the fixed-effect model, but now apply random-effects weights. Note that this approach only works if we use the same weights to compute the overall effect as we do to compute the effects within groups. In Table 19.5, studies from subgroup *A* use the T^2 value of 0.0164 both for computing the subgroup mean and for computing the overall mean. Similarly, studies from subgroup *B* use the T^2 value of 0.0022 both for computing the subgroup mean and for computing the overall mean.

We compute the following quantities (where *SS* is the sum of squared deviations).

- Q^*_A, the weighted *SS* of all *A* studies about the mean of *A*.
- Q^*_B, the weighted *SS* of all *B* studies about the mean of *B*.
- Q^*_{within}, the sum of Q^*_A and Q^*_B.
- Q^*_{bet}, the weighted *SS* of the subgroup means about the grand mean.
- Q^*, the weighted *SS* of all effects about the grand mean.

We may write $Q^*_{within} = Q^*_A + Q^*_B$, to represent the sum of within-group weighted *SS*, or more generally, for *p* subgroups,

$$Q^*_{within} = \sum_{j=1}^{p} Q^*_j. \tag{19.29}$$

In the running example

$$Q^*_{within} = 3.3882 + 3.9523 = 7.3406. \tag{19.30}$$

The weighted *SS* are additive, such that $Q^* = Q^*_{within} + Q^*_{bet}$. Therefore, Q^*_{bet} can be computed as

$$Q^*_{bet} = Q^* - Q^*_{within}. \tag{19.31}$$

Under the null hypothesis that the effect sizes μ are the same for all groups, 1 to *p*, Q^*_{bet} would be distributed as chi-squared with degrees of freedom equal to $p - 1$.

In the running example

$$Q^*_{bet} = 15.3952 - 7.3406 = 8.0547. \tag{19.32}$$

Results are summarized in Table 19.7. Note that the only *Q* statistic that we interpret here is the one *between groups*. In the running example, the *Between* line tells us

Table 19.7 Random-effects model (separate estimates of τ^2) – ANOVA table.

	*Q**	*df*	*p*	Formula
A	3.3882			19.19
B	3.9523			19.20
Within	7.3406			19.30
Between	8.0547	1.0	0.0045	19.32
Total	15.3952			19.23

that the difference between groups (the combined effect for *A* versus *B*) is statistically significant ($Q^*_{bet} = 8.0547$, $df = 1$, $p = 0.0045$), which means that the effect size *is* related to the frequency of tutoring. In Excel, the function to compute a *p*-value for *Q* is = CHIDIST(*Q*,*df*). For the test of *A* versus *B*, =CHIDIST(8.0547,1) returns 0.0045.

To address the statistical significance of the total variance or the variance within groups, we use the statistics reported using the fixed-effect weights (see Table 19.3) rather than using Q^* (total), Q^*_A, Q^*_B or Q^*_{within}.

Comparing *A* versus *B*: a *Q*-test for heterogeneity (Method 3)

Finally, we could treat the subgroups as if they were studies and perform a test for heterogeneity across studies. If we extract the two subgroup lines and the total line from Figure 19.5 and replace the diamonds with squares we get Figure 19.7.

Concretely, we start with two *studies* with effect sizes of 0.324 and 0.610, and variances of 0.006 and 0.004. Then, we apply the usual meta-analysis methods to compute *Q*. Concretely, using the values in Table 19.8, and applying (11.2) and subsequent formulas, we compute

$$M^* = \frac{214.163}{424.323} = 0.5047, \tag{19.33}$$

and

$$V_{M^*} = \frac{1}{424.323} = 0.0024. \tag{19.34}$$

$$Q = 116.146 - \left(\frac{214.163^2}{424.323}\right) = 8.0547,$$

$$df = 2 - 1 = 1,$$

and

$$p(Q = 8.0547, df = 1) = 0.0045,$$

where *Q* represents the weighted sum of squares for Studies *A* and *B* about the grand mean. For $Q = 8.0547$ and $df = 1$, the *p*-value is 0.0045.

In Excel, the function to compute a *p*-value for *Q* is =CHIDIST(*Q*,*df*). For the test of *A* versus *B*, =CHIDIST(8.0547,1) returns 0.0045.

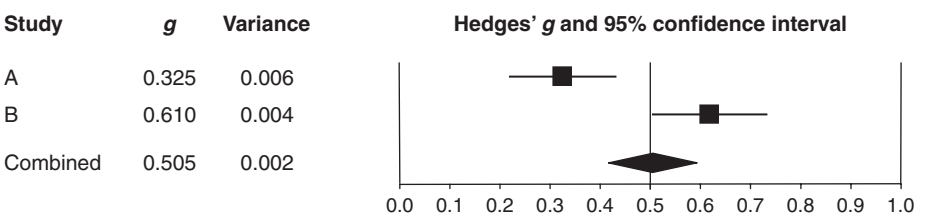

Figure 19.7 Random-effects model (separate estimates of τ^2) – treating subgroups as studies.

Table 19.8 Random-effects model (separate estimates of τ^2) – subgroups as studies.

Study	Effect size Y	Variance Within V_Y	Variance Between T^2	Variance Total V	Weight W	Calculated quantities		
						WY	WY^2	W^2
A	0.3245	0.0064	0.0000	0.0064	156.512	50.787	16.480	24495.944
B	0.6100	0.0037	0.0000	0.0037	267.811	163.376	99.666	71722.774
					424.323	214.163	116.146	96218.718

Quantify the magnitude of the difference

The difference and confidence interval are given by (19.17) and (19.18):

$$Diff^* = 0.6100 - 0.3245 = 0.2856,$$

$$SE_{Diff^*} = \sqrt{0.0064 + 0.0037} = 0.1006,$$

$$LL_{Diff^*} = 0.2856 - 1.96 \times 0.1006 = 0.0883,$$

and

$$UL_{Diff^*} = 0.2856 + 1.96 \times 0.1006 = 0.4828.$$

In words, the true difference between the effect in the *A* studies, as opposed to the *B* studies, probably falls in the range of 0.09 to 0.48.

RANDOM EFFECTS WITH POOLED ESTIMATE OF τ^2

Here, we show the computation of summary effects within subgroups, using a random-effects model with a pooled estimate of τ^2, which we refer to as τ^2_{within}. We illustrate the procedure in Figure 19.8. Note the common value of τ^2_{within} is assumed to apply to both subgroups.

Formula for estimating a pooled T^2

To estimate the pooled τ^2, proceed as follows. Recall (12.2) to (12.5) that to estimate τ^2 for a single collection of studies we use

$$T^2 = \frac{Q - df}{C}, \tag{19.35}$$

where

$$df = k - 1, \tag{19.36}$$

where k is the number of studies, and

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}. \tag{19.37}$$

In these equations, $Q - df$ is the excess (observed minus expected) sum of squared deviations from the weighted mean, and C is a scaling factor.

Similarly, to yield a pooled estimate of τ^2 we sum each element (Q, df, and C) across subgroups and then perform the same computation. Concretely,

$$T^2_{within} = \frac{\sum_{j=1}^{p} Q_j - \sum_{j=1}^{p} df_j}{\sum_{j=1}^{p} C_j}. \tag{19.38}$$

While the true value of τ^2_{within} cannot be less than zero (a variance cannot be negative), this method of estimating τ^2_{within} can yield a negative value due to sampling issues (when the observed dispersion is less than we would expect by chance). In this case, the estimate T^2_{within} is set to zero.

Computing the effects

Subgroup A yielded an estimate of 0.0164 while subgroup B yielded an estimate of 0.0122, represented in Figure 19.8 as T^2_A and T^2_B. We will pool these two estimates to yield a pooled value, represented as T^2_{within}, of 0.0097 (see (19.39)). This is the value used to assign weights in Table 19.10.

In the running example, the values within each group were computed earlier for A and B. Table 19.9 shows the values needed to calculate a pooled estimate T^2_{within} for the running example.

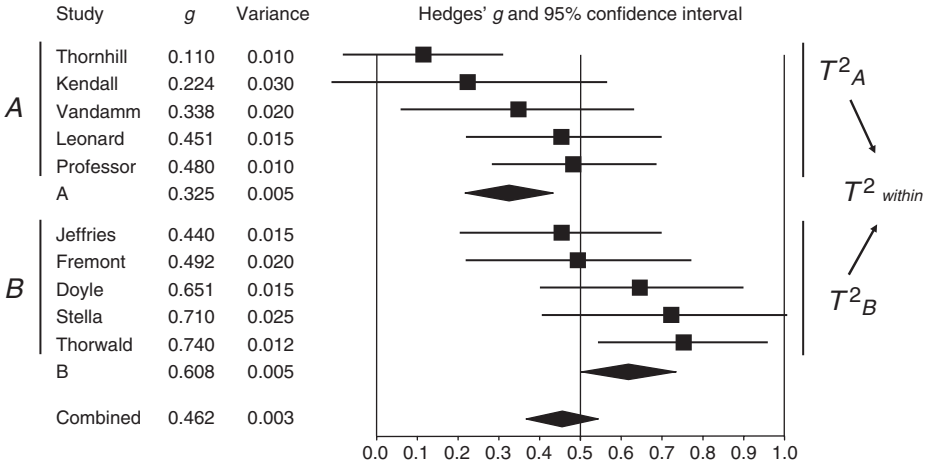

Figure 19.8 Random-effects model (pooled estimate of τ^2) – studies and subgroup effects.

Table 19.9 Statistics for computing a pooled estimate of τ^2.

Group	*Q*	*df*	*C*
A	8.4316	4	269.8413
B	4.5429	4	241.6667
Sum	12.9745	8	511.5079

Then,

$$T^2_{within} = \frac{12.9745 - 8}{511.508} = 0.00974. \tag{19.39}$$

Computations below are based on the values in Table 19.10. These are similar to Table 19.5, except that we now assume that all groups have the same τ^2, and use a common estimate. In Table 19.10 the same estimate of τ^2 (0.0097) is applied to all ten studies.

Table 19.10 Random-effects model (pooled estimate of τ^2) – computations.

	Study	Effect size *Y*	Variance Within V_Y	Variance Between T^2	Variance Total *V*	Weight *W*	Calculated quantities		
							WY	WY^2	W^2
A	Thornhill	0.110	0.0100	0.0097	0.0197	50.697	5.577	0.613	2570.150
A	Kendall	0.224	0.0300	0.0097	0.0397	25.173	5.639	1.263	633.678
A	Vandamm	0.338	0.0200	0.0097	0.0297	33.642	11.371	3.843	1131.752
A	Leonard	0.451	0.0150	0.0097	0.0247	40.445	18.241	8.226	1635.767
A	Professor	0.480	0.0100	0.0097	0.0197	50.697	24.334	11.681	2570.150
	Sum A					200.652	65.161	25.627	8541.498
B	Jefferies	0.440	0.0150	0.0097	0.0247	40.445	17.796	7.830	1635.767
B	Fremont	0.492	0.0200	0.0097	0.0297	33.642	16.552	8.143	1131.752
B	Doyle	0.651	0.0150	0.0097	0.0247	40.445	26.329	17.140	1635.767
B	Stella	0.710	0.0250	0.0097	0.0347	28.798	20.446	14.517	829.299
B	Thorwald	0.740	0.0120	0.0097	0.0217	46.030	34.062	25.206	2118.721
	Sum B					189.358	115.185	72.837	7351.306
	Sum					390.010	180.346	98.463	15892.804

Computations (random effects, pooled estimate of τ^2) for the A studies

$$M^*_A = \frac{65.161}{200.652} = 0.3247,$$

$$V_{M^*_A} = \frac{1}{200.652} = 0.0050,$$

$$SE_{M^*_A} = \sqrt{0.0050} = 0.0706,$$

$$LL_{M_A^*} = 0.3247 - 1.96 \times 0.0706 = 0.1864,$$

$$UL_{M_A^*} = 0.3247 + 1.96 \times 0.0706 = 0.4631,$$

$$Z_A^* = \frac{0.3247}{0.0706} = 4.6601,$$

$$p(Z_A^*) < 0.0001,$$

and

$$Q_A{}^* = 25.627 - \left(\frac{65.161^2}{200.652}\right) = 4.4660. \tag{19.40}$$

Note. The Q^* statistic computed here, using random-effects weights, is used *only* for the analysis of variance, to partition Q^* into its various components. Therefore, we do not show a p-value for Q^*. Rather, the Q statistic computed using fixed-effect weights (above) is the one that reflects the between-studies dispersion, provides a test of homogeneity for the studies within subgroup A, and is used to estimate τ^2_{within}.

Computations (random effects, pooled estimate of τ^2) for the B studies

$$M_B^* = \frac{115.185}{189.358} = 0.6083,$$

$$V_{M_B^*} = \frac{1}{189.358} = 0.0053,$$

$$SE_{M_B^*} = \sqrt{0.0053} = 0.0727,$$

$$LL_{M_B^*} = 0.6083 - 1.96 \times 0.0727 = 0.4659,$$

$$UL_{M_B^*} = 0.6083 + 1.96 \times 0.0727 = 0.7507,$$

$$Z_B^* = \frac{0.6083}{0.0727} = 8.3705,$$

$$p(Z_B{}^*) < 0.0001,$$

and

$$Q_B^* = 72.837 - \left(\frac{115.185^2}{189.358}\right) = 2.7706. \tag{19.41}$$

Computations (random effects, pooled estimate of τ^2) for all ten studies

The statistics here are computed using the same value of T^2 as was used within groups (in this case, the pooled estimate, T^2_{within}).

$$M^* = \frac{180.346}{390.010} = 0.4624, \tag{19.42}$$

$$V_{M^*} = \frac{1}{390.010} = 0.0026, \tag{19.43}$$

$$SE_{M^*} = \sqrt{0.0026} = 0.0506,$$

$$LL_{M^*} = 0.4624 - 1.96 \times 0.0506 = 0.3632,$$

$$UL_{M^*} = 0.4624 + 1.96 \times 0.0506 = 0.5617,$$

$$Z^* = \frac{0.4624}{0.0506} = 9.1321,$$

$$p(Z^*) < 0.0001,$$

and

$$Q^* = 98.463 - \left(\frac{180.346^2}{390.010}\right) = 15.0690. \tag{19.44}$$

The statistics computed above are summarized in Table 19.11.

Table 19.11 Random-effects model (pooled estimate of τ^2) – summary statistics.

	A	B	Combined
Y	0.3247	0.6083	0.4624
V	0.0050	0.0053	0.0026
SE_Y	0.0706	0.0727	0.0506
LL_Y	0.1864	0.4659	0.3632
UL_Y	0.4631	0.7507	0.5617
Z	4.6001	8.3705	9.1321
$p2$	0.0000	0.0000	0.0000
Q	4.4660	2.7706	15.0690

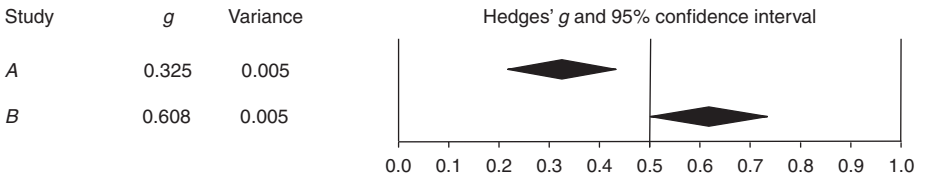

Figure 19.9 Random-effects model (pooled estimate of τ^2) – subgroup effects.

Comparing the effects

If we return to Figure 19.8 and excerpt the diamonds for the two subgroups we get Figure 19.9. The mean effect size for subgroups *A* and *B* are 0.325 and 0.608, with variances of 0.005 and 0.005.

Our goal is to compare these two mean effects, and there are several ways that we can proceed. These approaches are algebraically equivalent, and (it follows) yield the same *p*-value.

Comparing *A* versus *B*: a *Z*-test (Method 1)

We can use a *Z*-test to compare the mean effect for subgroups *A* versus *B*. The null hypothesis and formulas are the same as those for the prior case (where we did not assume a common value for τ^2). If we elect to subtract the mean of *A* from the mean of *B*,

$$Diff^* = M_B^* - M_A^*, \tag{19.45}$$

the test statistic to compare the two effects is

$$Z_{Diff}^* = \frac{Diff^*}{SE_{Diff^*}}, \tag{19.46}$$

where

$$SE_{Diff^*} = \sqrt{V_{M_A^*} + V_{M_B^*}}. \tag{19.47}$$

Under the null hypothesis that the true mean effect size μ_i is the same for both groups,

$$H_0 : \mu_A = \mu_B, \tag{19.48}$$

Z_{Diff}^* would follow the normal distribution. For a two-tailed test the *p*-value is given by

$$p^* = 2\left[1 - \left(\Phi\left(|Z_{Diff}^*|\right)\right)\right], \tag{19.49}$$

where $\Phi(Z)$ is the standard normal cumulative distribution.

In the running example

$$Diff^* = 0.6083 - 0.3247 = 0.2835,$$

$$SE_{Diff^*} = \sqrt{0.0050 + 0.0053} = 0.1013,$$

and

$$Z^*_{Diff} = \frac{0.2835}{0.1013} = 2.7986.$$

The two-tailed p-value corresponding to $Z^*_{Diff} = 2.7986$ is 0.0051. This tells us that the mean effect is probably not the same for the A studies as for the B studies. In Excel, the function to compute a 2-tailed p-value for Z is =(1-(NORMSDIST(ABS(Z))))*2. Here, =(1-(NORMSDIST(ABS(2.7986))))*2 will return the value 0.0045.

Comparing *A* with *B*: a *Q*-test based on analysis of variance (Method 2)

Again, we apply the same formulas as we did for the prior case, but this time using the random-effects weights based on a pooled estimate of τ^2. Note that this approach only works if we use the same weights to compute the overall effect as we do to compute the effects within groups. In Table 19.10 we used a T^2 value of 0.0097 for all ten studies, and this is the value used to sum *within* subgroups and also to sum *across* subgroups.

We compute the following quantities (where *SS* is the sum of squared deviations).

- Q^*_A, the weighted *SS* of all *A* studies about the mean of *A*.
- Q^*_B, the weighted *SS* of all *B* studies about the mean of *B*.
- $Q^*_{within,}$ the sum of Q^*_A and Q^*_B.
- $Q^*_{bet,}$ the weighted *SS* of the subgroup means about the grand mean.
- $Q^*_,$ the weighted *SS* of all effects about the grand mean.

We may write $Q^*_{within} = Q^*_A + Q^*_B$, to represent the sum of within-group weighted *SS*, or more generally, for p subgroups,

$$Q^*_{within} = \sum_{j=1}^{p} Q^*_j. \tag{19.50}$$

In the running example,

$$Q^*_{within} = 4.4660 + 2.7706 = 7.2366. \tag{19.51}$$

The weighted *SS* are additive, such that $Q^* = Q^*_{within} + Q^*_{bet}$. Therefore, Q^*_{bet} can be computed as

$$Q^*_{bet} = Q^* - Q^*_{within}. \tag{19.52}$$

Under the null hypothesis that the true man effect size μ is the same for all groups, 1 to p, Q^*_{bet} would be distributed as chi-squared with degrees of freedom equal to $p - 1$.

In the running example

$$Q^*_{bet} = 15.0690 - 7.2366 = 7.8324. \tag{19.53}$$

Table 19.12 Random-effects model (pooled estimate of τ^2) – ANOVA table.

	Q^*	*df*	*p*	Formula
A	4.4660			19.40
B	2.7706			19.41
Within	7.2366			19.51
Between	7.8324	1	0.0051	19.53
Total	15.0690			19.44

The only Q statistic that we interpret here is the one between groups. In the running example, the *Between* line tells us that the difference between groups (the combined effect for *A* versus *B*) is statistically significant ($Q^*_{bet} = 7.8324$ $df = 1$, $p = 0.0051$), which means that the effect size is related to the frequency of tutoring. In Excel, the function to compute a *p*-value for Q is =CHIDIST(Q,*df*). For the test of *A* versus *B*, =CHIDIST(7.8324,1) returns 0.0051.

To address the statistical significance of the total variance or the variance within groups, we use the statistics reported using the fixed-effect weights (Table 19.3) rather than using Q^*_{total}, Q^*_A, Q^*_B or Q^*_{within}.

Comparing *A* versus *B*: a *Q*-test for heterogeneity (Method 3)

Finally, we could treat the subgroups as if they were studies and perform a test for heterogeneity across studies. If we extract the two subgroup lines and the total line from Figure 19.8 and replace the diamonds with squares we obtain Figure 19.10.

Concretely, we start with two *studies* with effect sizes of 0.325 and 0.608, and variances of 0.005 and 0.005. Then, we apply the usual meta-analysis methods to compute Q. Concretely, using the values in Table 19.13, and applying (11.2) and subsequent formulas, we compute

$$M^* = \frac{180.346}{390.010} = 0.4624, \tag{19.54}$$

$$V_{M^*} = \frac{1}{390.010} = 0.0026, \tag{19.55}$$

$$Q = 91.227 - \left(\frac{180.346^2}{390.010}\right) = 7.8324,$$

$$df = 2 - 1 = 1,$$

and

$$p(Q = 7.8324, df = 1) = 0.0051.$$

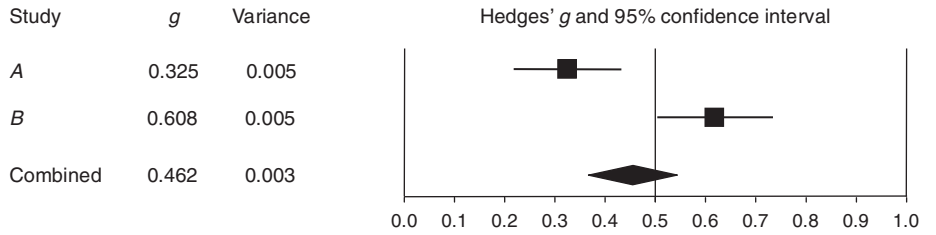

Figure 19.10 Random-effects model (pooled estimate of τ^2) – treating subgroups as studies.

Table 19.13 Random-effects model (pooled estimate of τ^2) – subgroups as studies.

Study	Effect size Y	Variance Within V_Y	Variance Between T^2	Variance Total V	Weight W	Calculated quantities WY	WY^2	W^2
A	0.3247	0.0050	0.0000	0.0050	200.652	65.161	21.161	40261.386
B	0.6083	0.0053	0.0000	0.0053	189.358	115.185	70.066	35856.405
					390.010	180.346	91.227	76117.791

where Q represents the weighted sum of squares for Studies A and B about the grand mean. For $Q = 7.8324$ and $df = 1$, the p-value is 0.0051.

In Excel, the function to compute a p-value for Q is =CHIDIST(Q,df). For example, =CHIDIST(7.8324,1) returns 0.0051.

Quantify the magnitude of the difference

The difference and confidence interval are given by (19.17) and (19.18):

$$Diff^* = 0.6083 - 0.3247 = 0.2835,$$

$$SE_{Diff^*} = \sqrt{0.0050 + 0.0053} = 0.1013,$$

$$LL_{Diff^*} = 0.2835 - 1.96 \times 0.1013 = 0.0850,$$

and

$$UL_{Diff^*} = 0.2835 + 1.96 \times 0.1013 = 0.4821.$$

In words, the true difference between the effect in the A studies, as opposed to the B studies, probably falls in the range of 0.09 to 0.48.

THE PROPORTION OF VARIANCE EXPLAINED

In primary studies, a common approach to describing the impact of a covariate is to report the proportion of variance explained by that covariate. That index, R^2, is defined as the ratio of explained variance to total variance,

$$R^2 = \frac{\sigma^2_{explained}}{\sigma^2_{total}} \tag{19.56}$$

or, equivalently,

$$R^2 = 1 - \left(\frac{\sigma^2_{unexplained}}{\sigma^2_{total}} \right). \tag{19.57}$$

This index is intuitive because it can be interpreted as a ratio, with a range of 0 to 1 (or expressed as a percentage in the range of 0% to 100%). Many researchers are familiar with this index, and have a sense of what proportion of variance is likely to be explained by different kinds of covariates or interventions.

This index cannot be applied directly to meta-analysis for the following reason. In a primary study, a covariate that explains all of the variation in the dependent variable will reduce the error to zero (and R^2, the proportion of variance explained, would reach 100%).

For example, Figure 19.11 depicts a primary study with 10 participants. All those in group *A* have the same score (0.3) and all those in group *B* have the same score (0.7). Since the variance *within* each subgroup is 0.0, group membership explains 100% of the original variance, and R^2 is 100%. In a real study, of course, there would be some variance within groups and R^2 would be less than 100%, but the fact that R^2 can potentially reach 100% is part of what makes this index intuitive.

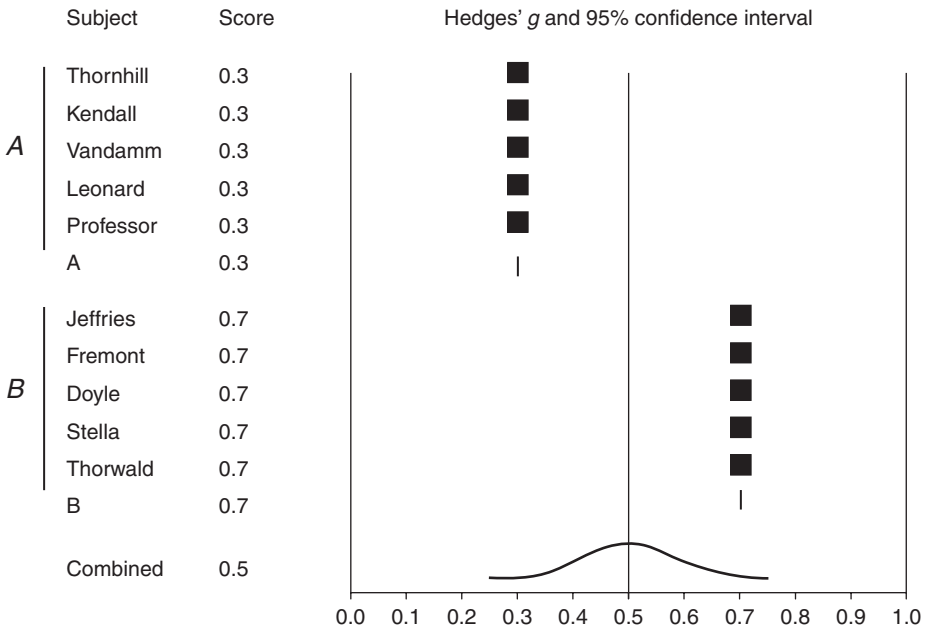

Figure 19.11 A primary study showing subjects within groups.

By contrast, consider what happens in a meta-analysis if we have two subgroups of studies. We assume that there are five studies in each subgroup, with a *true* summary effect (say, a standardized mean difference) of 0.30 for each study in subgroup *A* and of 0.70 for each study in subgroup *B*. However, while the true effect is identical for each study within its subgroup, the observed effects will differ from each other because of random error.

Thus, the variance within groups, while smaller than the variance between groups, can never approach zero. If the within-study error is a substantial portion of the total variance observed (say, 75%), then the upper limit of R^2 would be only 25%. As such, two important qualities of the index (the fact that it has a natural scale of 0% to 100% and the fact that it has the same range across studies) would no longer apply.

Since the problem with using R^2 is the fact that study-level covariates in a meta-analysis can address only the true variance τ^2 (and not the within-study variance v), the logical solution is to redefine R^2 (or to define a new index) that is based solely on the true variance. Rather than defining R^2 as the proportion of *total* variance explained by the covariates, we will define it as the proportion of *true* variance explained by the covariates. Since the true variance is estimated as T^2, this gives us

$$R^2 = \frac{T^2_{explained}}{T^2_{total}}, \tag{19.58}$$

or

$$R^2 = 1 - \left(\frac{T^2_{unexplained}}{T^2_{total}}\right). \tag{19.59}$$

In the context of subgroups, the numerator in (19.59) is the between-studies variance within subgroups, and the denominator is the total between-studies variance (within-subgroups plus between-subgroups). Therefore, the equation can be written

$$R^2 = 1 - \left(\frac{T^2_{within}}{T^2_{total}}\right). \tag{19.60}$$

In the running example, T^2 for the full set of studies was 0.0299 (see page 155), and T^2 computed by working within subgroups and then pooling across subgroups was 0.0097 (see page 173). This gives us

$$R^2 = 1 - \left(\frac{0.0097}{0.0299}\right) = 0.6745. \tag{19.61}$$

In Figure 19.12 we have superimposed a normal curve for the distribution of true effects within each subgroup of studies, and also across all ten studies. The relatively narrow dispersion within groups is based on the T^2 of 0.0097, while the relatively wide dispersion across groups is based on the T^2 of 0.0299, and R^2 captures this change.

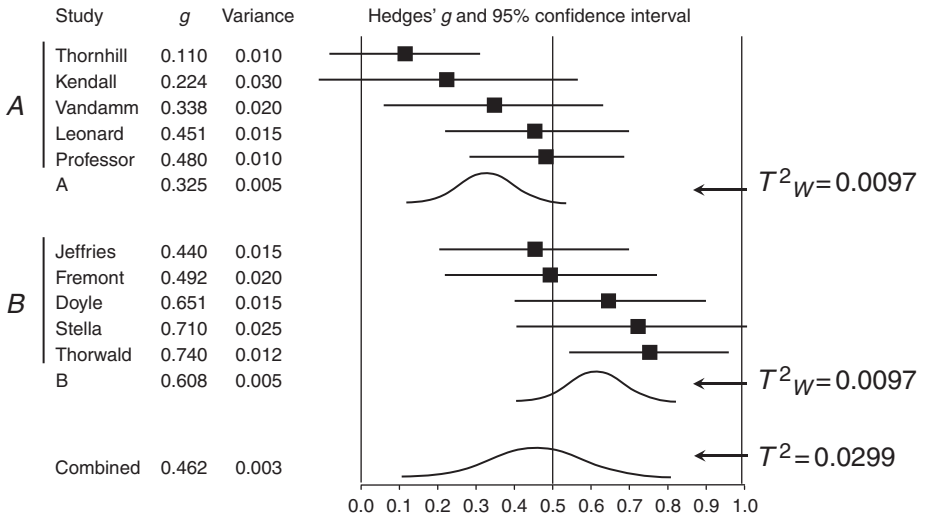

Figure 19.12 Random-effects model – variance within and between subgroups.

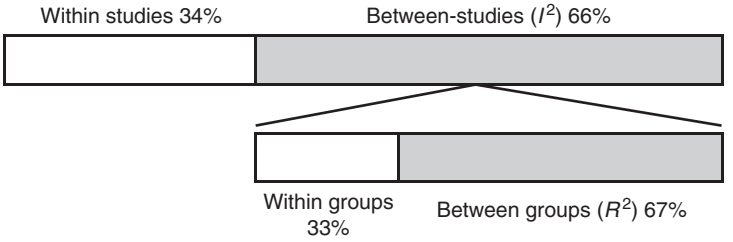

Figure 19.13 Proportion of variance explained by subgroup membership.

The same idea is shown from another perspective in Figure 19.13. On the top line, 34% of the total variance was within studies and 66% was between studies (which is also the definition of I^2). The within-studies variance cannot be explained by study-level covariates, and so is removed from the equation and we focus on the shaded part. On the bottom line, the type of intervention is able to explain 67% of the *relevant* variance, leaving 33% unexplained. Critically, the 67% and 33% sum to 100%, since we are concerned only with the variance between studies.

Note 1. While the R^2 index has a range of 0 to 1 (0% to 100%) in the population, it is possible for sampling error to yield an observed value of R^2 that falls outside of this range. In that case, the value is set to either 0 (0%) or 1 (100%).

Note 2. The R^2 index only makes sense if we are using a random-effects model, which allows us to think about explaining some of the between-studies variance. Under the fixed-effect model the between-studies variance is set at zero and cannot be changed. Also, the computational model proposed here for

estimating R^2 only works for the case where we assume that τ^2 is the same for all subgroups.

MIXED-EFFECTS MODEL

In this volume we have been using the term fixed effect to mean that the effect is identical (*fixed*) across all relevant studies (within the full population, or within a subgroup).

In fact the use of the term *fixed effect* in connection with meta-analysis is at odds with the usual meaning of *fixed effects* in statistics. A more suitable term for the fixed-effect meta-analysis might be a *common-effect* meta-analysis. The term *fixed effects* is traditionally used in another context with a different meaning. Concretely, we can talk about the subgroups as being *fixed* in the sense of fixed rather than random. For example, if we want to compare the treatment effect for a subgroup of studies that enrolled only males versus a subgroup of studies that enrolled only females, then we would assume that the subgroups are *fixed* in the sense that anyone who wanted to perform this analysis would need to use these same two subgroups (male and female). By contrast, if we have subgrouping of studies by country, then we might prefer to treat the subgroups as random. A random-effects assumption across subgroups of studies in the US, UK, Japan, Australia and Sweden would allow us to infer what the effect might be in a study in Israel, by assuming it comes from the same random-effects distribution. In this chapter we assume that when we are interested in comparing subgroups we make an assumption of the first type, which means that anyone who performs this comparison must use the same set of subgroups.

We mention this here for two reasons. One is to alert the reader that in the event that the subgroups have been sampled at random from a larger pool (as in the example of countries), then we are able to take this additional source of variability into account. The mechanism for doing so is beyond the scope of an introductory book.

The other reason is to explain the meaning of the term *mixed model*, which is sometimes used to describe subgroup analyses. As explained in this chapter the summary effect *within subgroups* can be computed using either a fixed-effect model or a random-effects model. As outlined immediately above, the difference *across subgroups* can be assessed using either a fixed-effects model or a random-effects model (although the meaning of *fixed* is different here). This leads to the following nomenclature.

If we use a fixed-effect model within subgroups and also across subgroups, the analysis is called a fixed-effects analysis. If we use a random-effects model within subgroups and a fixed-effect model across subgroups (the approach that we generally advocate), the model is called a mixed-effects model. We have the further possibility of assuming random effects both within and across subgroups; such a model is called a random-effects (or fully random-effects) model.

OBTAINING AN OVERALL EFFECT IN THE PRESENCE OF SUBGROUPS

In the tables and forest plots presented in this chapter we presented a summary effect for each subgroup and also for the total population. Since our primary concern has been with looking at difference between subgroups we paid little attention to the value for the total population. Here, we consider if that value should be reported at all, and if so, how it should be computed.

Should we report a summary effect across all subgroups?

The question of whether or not we should report a summary effect across all subgroups depends on our goals and also on the nature of the data.

Suppose the primary goal of the analysis is to see if a treatment is more effective among acute patients than among chronic patients, and it emerges that the treatment is very effective in one group but harmful in the other. In this case, the take-home message should be that we need to look at each group separately. To report that the treatment is moderately effective (on average) would be a bad idea since this is true for neither group and misrepresents the core finding. In this case, it would be better to report the effect for the separate subgroups only.

By contrast, if it turns out that the treatment is equally effective (or nearly so) in both subgroups, then it might be helpful to report a combined effect to serve as a summary. This would probably be the case also if there are minor differences among groups, but the substantive implication of the treatment (or the relationship) is the same for all groups. This is especially true if there are many subgroups, and the reader will be looking for a single number that is easy to recall.

If we do decide to report a combined effect across subgroups, we need to be clear about what this value represents, since this determines how it will be computed. The basic options are explained below.

Option 1. Combine subgroup means and ignore between-subgroup variance

One option is to compute the weighted mean of the subgroup means. In other words, we treat each subgroup as a study and perform a fixed-effect analysis using the mean effect and variance for each subgroup. In this chapter, we showed three versions of this approach.

These computations were shown for the fixed-effect model in (19.14) and (19.15) and where we computed the weighted mean of the two subgroups. Note that we would get the identical values if we worked with the original ten studies and weighted each by its fixed-effect weight (see (19.3) and (19.4)).

These computations were shown for the random-effects model with separate estimates of τ^2 in (19.33) and (19.34), where we computed the weighted mean of the two subgroups. Note that we would get the identical values if we worked with

the original ten studies and weighted each by its random-effects weight, with a separate estimate of τ^2 for each subgroup (see (19.21) and (19.22)).

These computations were shown for the random-effects model with a pooled estimate of τ^2 in (19.54) and (19.55), where we computed the weighted mean of the two subgroups. Note that we would get the identical values if we worked with the original ten studies and weighted each by its random-effects weight, with a pooled estimate of τ^2 (see (19.42) and (19.43)).

In all three cases, the combined effect refers to no actual population but is rather the average of two different populations. If the subgroups were male and female then the combined effect is the expected effect in a population that included both males and females (in the same proportions as in the subgroups). As always, the standard error of the mean speaks to the precision of the mean, and not to the dispersion of effects across subgroups (which is treated as zero).

Option 2. Combine subgroup means, and model the between-subgroup variance

A second option is to assume a random-effects model across subgroups. In other words, all the formulas and concepts discussed in Chapter 12 are applied here, except that the unit of analysis is the subgroup rather than the study. This would make sense if the subgroups have been sampled at random from a larger group of relevant subgroups. For example, we have the mean effect of a treatment in the US and in Australia, but we want to estimate what the mean effect of that treatment would be across all relevant countries.

In this case we need to address precisely the same kinds of issues we addressed when discussing heterogeneity in Chapter 12. First, we compute a measure of between-subgroups dispersion, T^2_{bet}. Then, we compute a weighted mean of the subgroups, where the weights are based on the within-subgroup error and the between-subgroups variance. To the extent that the subgroup means differ from each other, the standard error of the combined effect will be increased (but this additional error will be diminished as additional subgroups are added).

We can also focus on the dispersion itself (as in Chapters 16 and 17). For example, we can use the estimate of τ^2_{bet} to build a prediction interval that gives us the expected range of effect sizes for the next country (in our example) selected at random.

Option 3. Perform a separate random-effects analysis on the full set of studies.

If we want to report a combined effect across subgroups, then a third option is simply to perform a separate random-effects meta-analysis including all of the studies, and ignoring subgroup membership. Rather than estimate τ^2 within subgroups (as we did before) we estimate it across all studies, and so it will tend to be larger.

Comparing the options

When our primary goal is to assess differences among subgroups, and use an analysis of variance table as part of the process, the combined effects across subgroups are computed using option 1. This yields a set of internally consistent data.

If we really care about the combined effect across subgroups then options 2 and 3 are the more logical choices. If the subgroups really have been selected at random from a larger set, then option 2 allows us to model the different sources of error separately and obtain a better estimate of the true confidence interval for the combined effect (as well as discuss prediction intervals for a future subgroup), and is probably the better choice. This assumes, of course, that we have sufficient information to obtain a reasonably precise estimate of the variance among subgroups. By contrast, if the subgrouping is not of major importance, or if multiple different subgroupings of the studies are being considered, then option 3 is the more logical choice.

SUMMARY POINTS

- Just as we can use *t*-tests or analysis of variance in primary studies to assess the relationship between group membership and outcome, we can use analogs of these procedures in meta-analysis to assess the relationship between subgroup membership and effect size.
- We presented three methods that can be used to compare the mean effect across subgroups. To compare the mean effect in two groups we can use a *Z*-test. To compare the mean effect in two or more groups we can use analysis of variance (modified for use with subgroups) or the *Q*-test of homogeneity. All three procedures are mathematically equivalent.
- These analyses may be performed using either the fixed-effect or the random-effects model within groups, but in most cases the latter is appropriate.
- In primary studies we use R^2 to reflect the proportion of variance explained by group membership. An analogous index, which reflects the proportion of true variance explained by subgroup membership, can be used for meta-analysis.