Chapter 22
Meta-Analysis

'Fett's Law: Never replicate a successful experiment'

Content list

By the end of this chapter you will understand:

1. What meta-analysis is.
2. How it helps in confirming research findings.
3. How to undertake a meta-analysis.

Introduction

Meta-analysis has become an important research strategy as it enables researchers to combine the results of many pieces of research on a topic to determine whether the finding holds generally. This is better than trying to assume that the findings of a single study have global meaning.

What is meta-analysis?

Each strand of a rope contributes to the strength of the rope. But the rope is stronger than any individual strand. Similarly, when a particular finding is obtained repeatedly,

under a variety of conditions, we are strongly confident that there exists a general principle. The results of small localized individual studies, no matter how well conducted, are often insufficient to provide us with confident answers to questions of general importance. *Meta-analysis* allows us to compare or combine results across a set of similar studies. In the individual study, the units of analysis are the individual observations. In meta-analysis the units of analysis are the results of individual studies.

The term meta-analysis means '*an analysis of analysis*'. A particular topic may have been replicated in various ways, using, for example, differently sized samples, and conducted in different countries under different environmental, social and economic conditions. Sometimes results appear to be reasonably consistent; others less so. Meta-analysis enables a rigorous comparison to be made rather than a subjective 'eyeballing'. However, the technique relies on all relevant information being available for each of the examined studies. If some crucial factors like sample size and methodology are missing then comparison is not feasible.

Meta-analysis. *An objective and quantitative methodology for synthesizing previous studies and research on a particular topic into an overall finding.*

If you are asked to write a report for your manager on the economics of recycling waste paper, or on the relationship between air travel fares and ticket sales, or on the relationship between money supply and mortgage rates, you will search for information from a variety of sources, including in-house documents, the Internet, the national and local libraries, etc. The strategy is to read studies relevant to the topic you wish to investigate, summarize the findings, and then integrate the existing knowledge. From this you may conclude that a particular variable is of crucial importance, or that the relationships between particular variables are worthy of note. This is the standard literature survey where you draw essentially subjective conclusions based on your critical evaluation of the literature. You often use a 'voting method' as a crude index of where the balance of results lies.

This method is flawed and inexact because:

- you are unable to deal with the large number of studies on a topic, and so focus on a small subset of studies, often without describing how the subset was selected;
- you often cite the conclusions of previous reviews without examining those reviews critically;
- you are interested in a particular issue so you might not be inclined to give full weight to evidence that is contrary to your own desired outcome.

As a result, your subjective conclusion may not accurately reflect the actual strength of the relationship. You can reduce this possibility by adding a meta-analysis to your review. This allows you to collect, code, compare or combine results from different studies and interpret using statistical methods similar to those used in primary data analysis, facilitating statistically guided decisions about the strength of observed effects and the reliability of results

across a range of studies. The result is an integrated review of findings that is more objective and exact than a narrative review.

Meta-analysis. *This is a more efficient and effective way to summarize the results of large numbers of studies than subjective judgement or eyeballing.*

Examples of meta-analytic studies

1 The first ever meta-analysis study (Smith and Glass, 1977) synthesized the results of nearly 400 controlled evaluations of psychotherapy and counselling to determine whether psychotherapy 'works'. They coded and systematically analyzed each study for the kind of experimental and control treatments used and the results obtained. They were able to show that, on the average, the typical psychotherapy client was better off than 75% of the untreated 'control' individuals.
2 Rosenthal (1994) used meta-analysis to summarize the results of 345 studies on experimenter effects that occur when the participants in an experiment respond in ways that correspond to the expectancies of the experimenter. Rosenthal investigated this effect in eight areas where the effect had been studied (e.g. learning material; person perception; athletic performance) and the mean effect size was 0.70. This suggests a strong effect size so we can confidently state that the researchers' expectancies considerably influence participants' behaviour as a general principle.
3 Iaffaldano and Muchinsky (1985) found from their meta-analysis that overall there is only a slight relationship between workers' job satisfaction and the quality of their performance.
4 Jenkins (1986) tracked down 28 published studies measuring the impact of financial incentives on workplace performance. Only 57% of these found a positive effect on performance and the overall effect was minimal.
5 Mullen and Copper (1994) conducted a meta-analysis of 49 studies from various fields (e.g. industrial, sport, military, social) to determine the relationship between team cohesion and team success. They reported a positive but small relationship, and that specific task interaction does not serve as a moderator variable. They also reported that real groups exhibit significantly stronger cohesion-success effects than artificial groups, and sport teams exhibit even stronger effects than non-sport real groups. The strongest relationship between cohesion and group success is present in sport teams, followed by military groups and then non-military groups.
6 Gully *et al*. (2002) examined 67 studies to determine whether team efficacy (a team's belief that it can successfully perform a particular task), and team potency (a team's belief in its capabilities across a range of tasks), are positively related to team performance. Overall they were able to support the existence of a strong relationship.
7 Thorsteinson (2003) compared 38 studies involving over 51,000 employees of the job attitudes of full- and part-time workers. Overall a wide range of studies involving professional, non-professional and gender-based studies revealed no significant differences

between full- and part-time workers in respect of job satisfaction, organizational commitment and intention to leave.

8 Hosada *et al*. (2003) gathered together 27 studies on the effects of physical attractiveness on job related outcomes. Attractive individuals were found to fare better than unattractive individuals in terms of a number of outcomes. The weighted mean effect size, *d*, was 0.37 for all studies. In addition, professionals were as susceptible to the bias as were college students, and attractiveness was as important for men as for women,

9 Judge *et al*. (2004) conducted a meta-analysis of 96 studies that investigated the relationship between leadership qualities and intelligence. Results indicated that the corrected correlation between intelligence and leadership is 0.21. Overall, results suggest that the relationship between intelligence and leadership is considerably lower than previously thought.

10 De Dreu and Weingart (2003) investigated task versus relationship conflict in their effects on team performance and team member satisfaction using 30 studies and found that relationship conflict had negative effects on team performance and team satisfaction but, contrary to previous theorizing, task conflict also had deleterious affects on both …

Although meta-analysts use a number of advanced techniques, we will concentrate on some basic techniques that incorporate fundamental statistical procedures like 'significance', 'p' and 'r', discussed earlier in the book, to give you the flavour of what meta-analysis is. Students interested in a more comprehensive and advanced treatment should consult specialized accounts in a number of advanced texts (e.g. Rosenthal, 1991).

Conducting a meta-analysis

There are three stages to this:

1 Identify the relevant variables.
2 Locate relevant research.
3 Conduct the meta-analysis.

Stage 1 Identify the relevant variables

This sounds easy, but like defining a hypothesis and determining research questions, you must be specific and clear about what your real focus is. You cannot simply say, 'I want to do a meta-analysis on attitude change research'. As Rosenthal indicates (1984), the unit of analysis in meta-analysis is the impact of variable *x* on variable *y.* So you must limit yourself to conducting an evaluation of a much smaller segment, such as the effects of different types of feedback on job performance, or the effects of different levels of autonomy on group achievement of objectives in the service industry.

Stage 2 Locate the relevant research

This topic has already been dealt with in Chapter 4. However, one issue that is vital and potentially serious for the meta-analyst is the *file drawer problem.*

The file drawer problem

Because many journal editors are reluctant to accept 'non-significant' results, researchers' file drawers (their 'C' drives and memory sticks) may contain unpublished studies that failed to yield significant results. If there were a substantial number of such studies in the file drawers, the meta-analyst's evaluation of the overall significance level may be unduly optimistic. The file drawer phenomenon is potentially serious for meta-analysis because it produces a biased sample – a sample of only those results published because they reported statistically significant results. This bias inflates the probability of making a Type II error (concluding that a variable has an effect when it does not). Studies that failed to be published are not available to be included in the meta-analysis. Because meta-analytic techniques ultimately lead to a decision based on available statistical information, an allowance must be made for the file drawer phenomenon. There are two ways of dealing with the file drawer problem.

First, uncover those studies that never reach print by identifying as many researchers as possible in the research area you are researching. Then send each a questionnaire, asking if any unpublished research on the issue of interest exists. This may be impracticable in some topics as identification of researchers is difficult and non-response may be high anyway. A second but more practical approach, suggested by Rosenthal (1991), involves calculating the number of studies averaging null results (i.e. that did not reach significance) that would be required to push the significance level for all studies, retrieved and unretrieved combined, to the 'wrong' side of $p = .05$. If the overall significance computed on the basis of the retrieved studies can be brought down to the wrong side of p *(i.e.,* $p > .05$) by the addition of just a few null results, then the original estimate of p is clearly *not robust* (i.e., not resistant to the file drawer threat).

Table 22.1 illustrates such a calculation. It shows a table of 'tolerance' values in which the rows represent the number of retrieved studies and the columns represent three different levels of the average statistical significance of the retrieved studies. The intersection of any row and column shows the sum of retrieved and unretrieved combined, down to the level of being just barely 'nonsignificant' at $p > .05$. Suppose we meta-analyzed eight studies and found the average 'p' value to be .05. The 64 that is shown tells us that it will take an additional 56 unretrieved studies averaging null results to bring the original average $p = .05$ based on eight studies (i.e., $64 - 8 = 56$) down to $p > .05$.

As a rule of thumb, it has been suggested that we regard as robust any combined results for which the tolerance level reaches $5k + 10$, where k is the number of studies retrieved (Rosenthal, 1991). In our example of eight studies retrieved, this means that we will be satisfied that the original estimate of $p = 0.05$ is robust if we feel that there are fewer than an additional $5(8) + 10 = 50$ studies with null results squirrelled away in file drawers. Because Table 22.1 shows a tolerance for an additional 56 studies, it appears to us that the original estimate is indeed robust.

Stage 3 Conduct the meta-analysis

When you have located relevant literature, collected your data, and are reasonably certain that the file drawer phenomenon isn't an issue, you are ready to apply one of the many available meta-analytic statistical techniques. The heart of meta-analysis is the statistical combination of results across studies. Therefore, as well as recording the methodology, sample, design, hypotheses, conclusions, etc., you must record information particularly

Table 22.1 Tolerances for future null results as a function of the original average level of significance per study and the number of studies summarized

Number of studies summarized	Original average significance level		
	.05	.01	.001
1	1	2	4
2	4	8	15
3	9	18	32
4	16	32	57
5	25	50	89
6	36	72	128
7	49	98	173
8	64	128	226
9	81	162	286
10	100	200	353
15	225	450	795
20	400	800	1412
25	625	1250	2206
30	900	1800	3177
40	1600	3200	5648
50	2500	5000	8824

Note: Entries in this table are the total number of old and new studies required to bring an original average p of .05, .01, or .001 down to $p > .05$ (i.e., 'non-significance').

from the results section of research papers you are reviewing such a r's, t's, chi squares, F's, and p values. Table 22.2 illustrates meta-analytic techniques that can be applied for simplicity of explanation to situations where you have two studies.

1 The first general technique is that of *comparing studies*. This comparison is made when you want to determine whether two studies produce significantly different effects.
2 The second general technique involves *combining studies* to determine the average effect size of a variable across studies.

For each approach, you can evaluate studies by comparing or combining either p-values or effect sizes.

Comparison of effect sizes of two studies is generally more desirable than simply looking at p values because effect sizes provide a better estimate of the degree of impact of a variable than does the p value. (Remember, all the p value tells you is the likelihood of making a Type I error.) The p values are used when the information needed to analyse effect sizes

Table 22.2 Meta-analytic techniques for comparing and combining two studies

Technique	Method/Purpose
Comparing Studies Used to determine if two studies produce significantly different results.	Significance Testing. Record p-values from research and convert them to exact p-values (such as a finding reported at $p < 0.05$ may actually be $p = .036$). Used when information is not available to allow for evaluation of effect sizes.
	Effect-Size Estimation. Record values of inferential statistics (*F*, *t*, for example), along with associated degrees of freedom. Estimate effect sizes from these statistics. Preferred over significance testing.
Combining Studies Used when you want to determine the potency of a variable across studies.	Significance Testing. Can be used after comparing studies to arrive at an overall estimate of the probability of obtaining the two p-values under the null hypothesis.
	Effect-Size Estimation. Can be used after comparing studies to evaluate the average impact across studies of an independent variable on the dependent variable

Based on Rosenthal (1984).

is not included in the studies reviewed. Consequently, the following discussion focuses on meta-analytic techniques that look at effect sizes. For simplicity, a case involving only two studies is discussed. The techniques discussed here can be easily modified for the situation where you have three or more studies. For more information, see Rosenthal (1979, 1984) and Mullen and Rosenthal (1985).

Replication and meta-analysis

Table 22.3 illustrates why researchers prefer to work with effect sizes rather than *p* values. Set A shows two results, with the *p* values both rejecting the null (i.e., both *p*'s = .05) and with a difference in effect sizes of .30 in units of r (i.e. .50 –.20 = .30). The fact that both studies were able to reject the null and at exactly the same *p* level is a function of sample size, whereas the difference in effect sizes implies the degree of failure to replicate. Set B shows two studies with different *p* values, one significant at $p < .05$ and the other not significant; the two effect sizes, on the other hand, are in excellent agreement. The meta-analyst would say, accordingly, that Set B shows more successful replication than does Set A. Set C shows two studies differing markedly in both level of significance and magnitude (and direction) of effect size. Observe that one of the effects is reported as a 'negative' r, which tells us that this result was not in the same direction as the other result. Set C, then, is a not very subtle example of a clear failure to replicate. That the combined probabilities of all three sets are identical to one another (combined $p = .0028$) tells us that the pooled significance level is uninformative in differentiating successful from unsuccessful sets of replication studies.

Table 22.3 Comparison of three sets of replications

	Set A		Set B		Set C	
	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
N	96	15	98	27	12	32
p level (two-tailed)	.05	.05	01	.18	.000001	.33
r (effect size index)	.20	.50	.26	.26	.72	–.18
Pooled p, i.e. combining both studies (one-tailed)	.0028		.0028		.0028	

Comparing studies by effect size

What is effect size?
In Chapter 11, we indicated that this is the *degree* to which the phenomenon is present in the population. In meta-analysis, the finding of a study is converted into an effect size estimate. Various ways of estimating effect size such as the standardized mean difference (d), the correlation coefficient (r), and *eta* are used. You should refer back to Chapter 11 for the various formulae used to calculate these estimates of effect size.

But before synthesizing the effect sizes of separate studies, the meta-analyst usually finds it instructive to compare the results to discover the degree of their actual similarity. One approach is to define effect size 'r's as dissimilar if they are significantly different from one another, and to define them as similar if they are not significantly different from one another. In statistical terms, such significance testing involves:

(1) converting the quoted statistic from both studies, e.g. 't' or chi square into 'r's.
(2) giving the calculated 'r's the same sign if both studies show effects in the same direction, but different signs if the results are in the opposite direction;
(3) finding for each 'r' the associated '*Fisher z*' value. Fisher's z (i.e. lowercase z to differentiate this statistic from the uppercase Z denoting the normal distribution measure, or Z score described in Chapter 8) refers to a set of log transformations of 'r', as shown in Table 22.4.
(4) substituting in the following formula to find the Z score:

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

Example
Imagine you are interested in comparing two experiments that investigated the impact of the credibility of a communicator on persuasion for similarity of effect size to determine

Table 22.4 Transformations of r into Fisher's z

r	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.000	.010	.020	.030	.040	.050	.060	.070	.080	.090
.1	.100	.110	.121	.131	.141	.151	.161	.172	.182	.192
.2	.203	.213	.224	.234	.245	.255	.266	.277	.288	.299
.3	.310	.321	.332	.343	.354	.365	.377	.388	.400	.412
.4	.424	.436	.448	.460	.472	.485	.497	.510	.523	.536
.5	.549	.563	.576	.590	.604	.618	.633	.648	.662	.678
.6	.693	.709	.725	.741	.758	.775	.793	.811	.829	.848
.7	.867	.887	.908	.929	.950	.973	.996	1.020	1.045	1.071
.8	1.099	1.127	1.157	1.188	1.221	1.256	1.293	1.333	1.376	1.422

whether it is worthwhile combining them. In the results sections of the two studies, you found the following information concerning the effect of credibility on persuasion:

Study 1: $t = 2.57$, $p < .01$, $N = 22$;
Study 2: $t = 2.21$, $p < .05$, $N = 42$.

The first thing you must do is to determine the size of the effect of communicator credibility in both studies. Unfortunately, neither study provides that information (you will rarely find such information). Consequently, you must estimate the effect size based on the available statistical information which is t. Using the formula for t (see Chapter 11) gives the following results:

Study 1: $r = \sqrt{6.59/(6.59+20)} = .50$
Study 2: $r = \sqrt{4.89/(4.89+40)} = .33$

The next step in the analysis is to convert the r-values into Fisher z-scores. This is necessary because the distribution of 'r' becomes skewed as the population value of r deviates from zero. Converting r to Fisher's z corrects for this skew (Rosenthal, 1984). Table 22.4 shows for the r-values calculated above, the Fisher's z-values are 0.55 and 0.34, respectively.

Now you test for the difference between the two Fisher z scores with the following formula:

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

In this example, you have:

$$Z = \frac{.55 - .34}{\sqrt{\frac{1}{19} - \frac{1}{39}}} = \frac{.21}{.279} = .75$$

This Z score (standard deviate) of .75 is then evaluated for statistical significance by using the areas under the normal curve (Chapter 8: see Table 22.5 below). We know that .75 is not significant beyond $p < .05$ (we need a value of 1.96 remember). Consequently, you would conclude that the effect sizes produced by the two evaluated studies do not differ significantly. They are therefore good candidates for combining. Had a significant difference been obtained ($Z > 1.96$) you would conclude that one study produced a significantly larger effect than the other study.

If you find a significant difference between effect sizes, you should investigate why the difference exists. You might look at the methods, materials, sample sizes and procedures used in each study, as any or all of these may differ considerably between the studies and may be likely causes of the different effects.

Combining studies by effect size

Given two effect sizes that are not significantly different and therefore combinable on statistical and/or logical grounds, you may want to determine the average size of an effect across studies. The formula to be used again employs the Fisher z transformation:

$$\text{Mean z or } z_m = \frac{z_1 + z_2}{2}$$

in which the denominator is the number of Fisher z scores in the numerator; the resulting value is an average (or z_m).

The first step to take when combining the effect sizes of two studies is to calculate 'r' for each and convert each r-value into corresponding z-scores. Using the data from the example above to demonstrate comparing studies, we already have z values of .55 and .34:

$$z_m = \frac{(.55 + .34)}{2} = .45$$

This z is reconverted to an 'r' using Table 22.4. The r-value associated with this average Fisher z is .42. Hence, you now know that the average effect size across these two studies is .42.

Example
Given effect size estimates of 'r' = .7 (N = 20) for study A and 'r' = .5 (N = 80) for study B, find a combined estimate of the effect size.

We first convert each 'r' into z scores and then substitute into the formula. This gives:

$$z_m = \frac{(.867 + .549)}{2} = .708$$

This average Fisher z converts back to a combined effect of 'r' = .65. This is larger than the mean of the two 'r's.

Table 22.5 Fractional parts of the total area (taken as 10,000) under the Normal Probability Curve, corresponding to distances on the baseline between the mean and successive points laid off the mean in units of Standard Deviation. Example: Between the mean, and a point 1.3, is found 40.32% of the entire area under the curve, or there is a probability of 4032 of a value occurring between 0 and 1.3Z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0.7	2580	2611	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3290	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4383	4394	4406	4418	4429	4441
1.6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2.0	4772	4780	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4855	4857
2.2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2.9	4981	4982	4982	4983	4984	4985	4985	4986	4986	
3.0	4986.5		4987.4		4988.2		4988.9		4989.7	
3.1	4990.3		4991.0		4991.6		4992.1		4992.6	
3.2	4993.129									
3.3	4995.166									
3.4	4996.631									
3.5	4997.674									
3.6	4998.409									

Continued

Table 22.5 —cont'd

z	.00	.01	. 02	.03	.04	.05	.06	.07	.08	.09
3.7	4998.922									
3.8	4999.277									
3.9	4999.519									
4.0	4999.683									
4.5	4999.966									
5.0	4999.997133									

Remember, always compare the studies before combining them. If the effect sizes of the two studies are statistically different, it makes little sense to average their effect sizes. If the results from the studies are in opposite directions (for example, one study shows a positive effect of the independent variable, while the second shows a negative effect) combining should never be considered.

Unfortunately, there is no established criterion to judge whether or not the combined effect size is significant or, for that matter, important.

Comparing studies by significance levels

Although meta-analysts are usually more interested in effect sizes (usually 'r's as we have seen) than *p* values, they sometimes evaluate the overall level of significance as a way of increasing power. It is again instructive to find out whether the individual values are homogeneous (i.e. telling the same story) and therefore combinable. To make such a comparison, the meta-analyst first needs an accurate *p* level quoted in the research paper – such as *p* = .43 or .024 or .0012.

For each p value the meta-analyst then finds Z (i.e. not the Fisher z, but the standard normal deviate Z) using the table of Z (Table 22.5). Both *p* values should also be one-tailed, and we give the corresponding Z's the same sign if both studies showed effects in the same direction, but different signs if the results are in the opposite direction. The difference between the two Z's when divided by $\sqrt{2}$ yields a new Z. This new Z corresponds to the *p* value of the difference between the Z's if the null hypothesis were true (i.e. if the two Z's did not really differ).

Recapping,

$$Z = \frac{Z_1 - Z_2}{\sqrt{2}}$$

is distributed as Z, so we can enter this newly calculated Z in a table of standard normal deviates to find the *p* value associated with a Z of the size obtained or larger.

Example
Suppose that Studies A and B yield results in opposite directions, and neither is 'significant'. One *p* is .075 one-tailed and the other *p* is .109 one-tailed, but in the opposite tail. The Z's corresponding to these *p* values, found in Table 22.5, are +1.44 and –1.23 (note the opposite signs which indicate results in opposite directions). To find the Z scores, subtract the *p* value away from .5000 as the Z table covers only half the normal distribution. For example, a *p* of .075 subtracted from .5000 =.4250. Looking in the body of the table the closest we get is .4251 that gives a Z = 1.44. Our equation is then:

$$Z = \frac{Z_1 - Z_2}{\sqrt{2}} = \frac{1.44 - (-1.23)}{\sqrt{2}} = \frac{2.67}{1.41} = .189$$

The *p* value (Table 22.5) associated with a Z of 0.189 is .0294 one-tailed (rounded to .03). The two p values may thus be seen to differ significantly (or nearly so, if we used the two-tailed *p* of .0294 × 2 = .0588), suggesting that the results in terms of the *p* values of the two studies are heterogeneous even when we allow for normal sampling fluctuations. Thus, the *p* levels should not be combined.

Here is another example. Imagine the *p*-values (one-tailed) for Study A and B are: *p* = .02 (significant), *p* = .07 (not significant). Compute new Z as:

$$Z = \frac{Z_1 - Z_2}{\sqrt{2}} = \frac{2.06 - 1.48}{1.41} = .41$$

Its *p* value is .341 (one-tailed) or .682 (two-tailed).

Hence, the difference between these two studies (one significant, the other not significant) is not significant.

Combining studies by significance levels

After we compare the results of two separate studies, it is an easy matter to combine the p levels. In this way, we get an overall estimate of the probability that the two *p* levels might have been obtained if the null hypothesis of no relation between X and Y were true. To perform these calculations, we modify the numerator of the formula for comparing p values that we just described. We obtain accurate *p* levels for each of our two studies and then find the Z corresponding to each of these *p* levels. Also as before, both *p*'s must be given in one-tailed form, and the corresponding Z's will have the same sign if both studies show effects in the same direction and will have different signs if the results are in the opposite direction.

The only change in the previous equation is to add the Z values instead of subtracting them:

$$Z = \frac{Z_1 + Z_2}{\sqrt{2}}$$

This new Z corresponds to the *p* value of the two studies combined if the null hypothesis of no relation between X and Y were true.

Example
Suppose studies A and B yield homogeneous results in the same direction but neither is significant. One *p* is .121, and the other is .084; their associated Z's are 1.17 and 1.38, respectively. From the preceding equation we have:

$$Z = \frac{Z_1 - Z_2}{\sqrt{2}} = \frac{1.17 + 1.38}{\sqrt{2}} = \frac{2.55}{1.41} = 1.81$$

as our combined Z. The *p* associated with Z = 1.81 is .035 one-tailed (or .07 two-tailed). This is significant one-tailed even though the original *p*'s were not.

Example
As another example, imagine *p* values (one-tailed) for Study A and B are: *p* =.02 (significant), *p* = .07 (not significant). The two *p*-values can be combined to obtain an estimate of the probability that the two *p* values might have been obtained if the null hypothesis of no relation between X and Y were true:

$$Z = \frac{Z_1 - Z_2}{\sqrt{2}} = \frac{2.06 + 1.48}{1.41} = 2.51$$

Its *p*-value is .006 (one-tailed) or .012 (two-tailed). This combined *p* value is significant; it supports the significant study A.

Comparing and combining more than two effect sizes and significance levels

Although the discussion in this chapter has focused on comparing or combining only two studies, you will probably want to compare more than two studies. The mathematical formulas used to meta-analyse several studies are a bit more complex than those used in the two-study case. However, the general logic applied to the two-study case applies to the multi-study case. The formulas for comparing and combining more than two studies can be found in Rosenthal (1984). A list of several well-used computer programs that do most of the tedious calculations needed to perform such meta-analyses are provided at the end of the chapter.

Some issues in meta-analysis

Meta-analysis can be a powerful tool to evaluate results across studies. Even though many researchers have embraced the concept of meta-analysis, others question its usefulness on several grounds. This section explores some of the drawbacks to meta-analysis and presents some of the solutions suggested to overcome those drawbacks.

Table 22.6 Summary of general procedures for meta-analysis

Step	Procedure	Caution
Identify and collect studies	Define the criteria used to include or exclude studies. Describe how the search is conducted and the studies collected. Search for unpublished studies to test for Type I error publication bias.	Tendency to include studies not very similar; mixing 'apples' and 'oranges'. Time consuming to locate unpublished studies. May have included studies of poor quality. Stronger effects are found in journal articles than in theses; unpublished studies tend to be non-significant – sampling bias.
Quantify criteria	Convert reported results to common metrics for meta-analysis, e.g., effect size, 't', 'r', 'p'.	Over-emphasis on a single value, i.e., effect size. A wide variety of effect size estimates and corrections. Effect size estimates not directly comparable among some studies due to arbitrary scales used.
Code characteristics of studies – the crucial aspect	Code substantive characteristics, e.g., nature of samples, types of instruction, classification of outcomes using theory. Code methodology characteristics, e.g., dropout rate, design used, source of studies, date of study. Check validity and reliability of coding.	No systematic or logical procedure to build coding. Consult the literature and others with respect to the coding used.
Analyse data	Average effect sizes. Estimate variation in effect sizes. Divide studies into subgroups and test for homogeneity.	Calculate parametric and non-parametric estimates, if possible.
Discuss results	Describe limitations of review. Provide guidelines for future research.	

Assessing the quality of the research reviewed. Not all journals are equally reliable sources. The quality of the research found in a journal depends on its editorial policy. Some journals have more rigorous publication standards than others. This means that the quality of published research may vary considerably from journal to journal. One problem facing the meta-analyst is how to deal with this uneven research quality. For example, should an article published in a non-refereed journal be given as much consideration as an article published in a well regarded reputable refereed journal? There is no simple answer.

While Rosenthal has suggested weighting articles according to quality, on what bases should they be weighted? The refereed/non-refereed dimension is one possibility. However, simply because an article was not refereed is not a reliable indicator of its quality. Research in a new area, using new methods, is sometimes rejected from refereed journals even though it is methodologically sound and of high quality. Conversely, publication in a refereed journal is no guarantee that the research is of high quality.

A second dimension along which research could be weighted is according to the soundness of methodology, regardless of journal quality. Several experts could rate each study for its quality (perhaps on a zero to 10 scale). The ratings would then be checked for inter-rater reliability and used to weight the degree of contribution of each study to the meta-analysis.

Combining/comparing studies using different methods. A frequent criticism of meta-analysis is that it is difficult to understand how studies with widely varying materials, measures, and methods can be compared. This is commonly referred to as the 'apples versus oranges argument' in that all we end up with is a statistical fruit salad.

However, comparing results from different studies is no different from averaging across heterogeneous subjects in an ordinary study. If you are willing to accept averaging across subjects, you should also be willing to accept averaging across heterogeneous studies. The core issue is not whether averaging should be done across heterogeneous studies, but rather whether or not differing methods are related to different effect sizes. If methodological differences appear to be related to the outcome of research, studies in a meta-analysis could be grouped by methodology to determine its effects.

Practical problems. The task facing a meta-analyst is a formidable one. Not only may studies on the same issue use widely different methods and statistical techniques, some studies may not provide the necessary information to conduct a meta-analysis and have to be eliminated. The problem of insufficient or imprecise information (along with the file drawer problem) may result in a non-representative sample of research being included in a meta analysis. Admittedly, the bias may be small, but nevertheless exist.

Do the results of meta-analysis differ from those of traditional reviews? A valid question is whether or not traditional literature reviews produce results that differ qualitatively from those of a meta-analysis. To answer this question, Cooper and Rosenthal (1980) directly compared the two methods. Graduate students and professors were randomly assigned to conduct either a meta-analysis or a traditional review of seven articles dealing with the impact of gender of subject on persistence in a task. Two of the studies showed that females were more persistent than males, whereas the other five either presented no statistical data or showed no significant effect.

The results of this study showed that subjects using the meta-analysis were more likely to conclude that there was an effect of sex on persistence than were subjects using the traditional method. Additionally, subjects doing the traditional review believed that the effect of sex on persistence was smaller than did subjects doing the meta-analysis. Overall, 68% of the meta-analysts were prepared to conclude that sex had an effect on persistence, whereas only 27% of subjects using the traditional method were so inclined.

What you have learned in this chapter

You have been introduced to meta-analysis as a quantitative tool for comparing or combining results across a set of similar studies, facilitating statistically guided decisions about the strength of observed effects and the reliability of results across a range of studies. Meta analysis is a more efficient and effective way to summarize the results of large numbers of studies.

- The first general technique is that of *comparing studies*. This comparison is made when you want to determine whether two studies produce significantly different effects.
- The second general technique involves *combining studies* to determine the average effect size of a variable across studies.

For each approach, you can evaluate studies by comparing or combining either *p*-values or effect sizes.

The problem posed by the file drawer phenomenon is potentially serious for meta-analysis because it results in a biased sample – a sample of only those results published because they produced acceptable statistically significant results. But even published research may be of uneven quality.

Review questions

Qu. 22.1

Suppose you have used 100 subjects to replicate a study that reported a large effect of r = 0.50 based on only 10 subjects. You find a smaller sized effect in yours of r = –0.31:

(a) do you code your effect as positive or negative?
(b) what are Fisher's z for each effect?
(c) compute Z and find the associated *p* value
(d) what are your conclusions about the merit of combining these two studies?

Qu. 22.2

In the above study in Qu. 22.1 suppose your result is r = 0.40 and N = 120 while the original study produced r = .45 with N = 120, find:

(a) the corresponding Fisher's z's.
(b) Z and p
(c) Comment on the implications for your results

Evaluate your results in the light of the material above

Qu. 22.3

Given two similar studies with effect sizes r = .45 and r = .40, both coded positive to show that the results were in the predicted direction, find:

(a) Fisher's z for each
(b) compute mean z and
(c) find combined effect size

Qu. 22.4
Suppose studies L and M yield results in the same direction. But only L is significant (p =.05) while M is not significant with $p = 0.07$, find:

(a) Z's corresponding to these *p*'s
(b) the difference between the Z's
(c) the *p* value associated with the new Z of the differences and comment on this p value

Review your answer from the material above.

Check the answers on this chapter's Web page.

References

Cooper, H. & Rosenthal, R. 1982. Statistical versus traditional methods for summarising research findings. *Psychological Bulletin*, 87, 442–449.

De Dreu, C.K. & Weingart, L.R. 2003. Task versus relationship conflict, team performance and team member satisfaction: a meta analysis. *Journal of Applied Psychology*, 88 (4), 741–749.

Gully, S., Incalcaterra, K., Joshi, A. & Beaubien, J. 2002. A meta analysis of team efficiency, potency and performance. *Journal of Applied Psychology*, 87 (5), 819–832.

Hosada, M., Stone-Romero, E. & Coats, G. 2003. The effects of physical attractiveness on job related outcomes. *Personnel Psychology*, 56 (2), 431–448.

Iaffaldano, M. & Muchinsky, P.M. 1985. Job satisfaction and job performance: a meta analysis. *Psychological Bulletin*, 97, 251–273.

Jenkins, J. 1986. Financial incentives. *In Generalising from Laboratory to Field Settings*. Locke, E. (ed). Lexington: Lexington Books.

Judge, T.A., Colberet, A. & Illies, R. 2004. Intelligence and leadership. *Journal of Applied Psychology*, 89 (3), 542–552.

Mullen, B. & Copper, C. 1994. The relation between group cohesion and performance. *Psychological Bulletin*, 115 (2), 210–227.

Mullen, B. & Rosenthal, R. 1985. *Basic Meta-analysis*. Hillsdale: Lawrence Erlbaum.

Rosenthal, R. 1979. The file drawer problem. *Psychological Bulletin,* 86, 638–641.

Rosenthal, R. 1984. Meta analytic procedures for social research, in *Applied Social Science Research Methods, Vol 6*. Beverly Hills: Sage.

Rosenthal, R. 1994. Interpersonal expectancy effects. A 30 year perspective. *Current Directions in Psychological Science*, 3, 176–179.

Rosenthal, R. 1991. *Meta-analytic Procedures for Social Research*. Newbury Park: Sage.

Smith, M. & Glass, G. 1977. Meta-analysis of psychotherapy outcome studies. *American Psychologist*. 32, 752–760.

Thorsteinson, T.J. 2003. Job attitudes of part-time vs. full-time workers. *Journal of Occupational and Organisational Psychology*, 76 (2), 151–177.

You will never have to undertake meta-analysis by hand. There are computer programs for this. For example:

1 *Meta-Stat – A Tool for the Meta-Analysis of Research Studies*. Produced by Lawrence M. Rudner, Gene V. Glass, David L. Evartt and Patrick J. Emery. Meta-Stat is a DOS-based computer program that automates the many complex tasks that are required to

perform a meta-analysis. The data can easily be output in a format ready for use by SPSS. Meta-Stat is free for non-commercial, educational use. Meta-Stat is available through the auspices of the ERIC Clearinghouse on Assessment and Evaluation, Department of Measurement, Statistics and Evaluation, University of Maryland, College Park.

2 *Comprehensive Meta-Analysis Program (CMA)* produced by Biostat. A free trial download is available at www.power.analysis.com/about.biostat.htm

Now turn to the website page for this chapter and undertake the activities there.