Meta-analysis. Online Supplement 1 to Serious stats: A guide to advanced statistics for the behavioral sciences. Basingstoke...

Chapter · January 2012

CITATIONS
0

READS
79

1 author:

Thom S Baguley
Nottingham Trent University
64 PUBLICATIONS **889** CITATIONS

SEE PROFILE

Online Supplement 1
Meta-analysis

This supplement draws primarily on Chapters 5, 7 and 9.

OS1.1 Combining effect sizes using meta-analysis

The literal meaning of meta-analysis is the analysis of other analyses. The term is sometimes broadly applied to research synthesis: systematic reviews of research involving a large set of studies.[1] Meta-analysis, in a narrow sense, refers to a formal statistical model of research findings. These findings can take different forms (e.g., *p* values), though most meta-analyses combine findings in the form of effect size statistics. This requires taking into account the magnitude and variability of each effect. There are many ways to do this. Arguably the best method is to combine the raw data from each analysis in a pooled analysis (e.g., in the form of a *multilevel regression model* – see Hox, 2010). Meta-analysis is particularly useful for combining findings from many small studies in one large analysis. This often involves data from many different studies (sometimes many years old), and it is rare that raw data is available for every study. For this reason, most meta-analyses involve combining effect size statistics reported in published work or derived from summary statistics. The best-known meta-analytic models employ odds ratios or standardized effect size metrics such as *r* or *g* (Field and Gillett, 2010; Hunter and Schmidt, 2004).

Shadish and Haddock (1994) provide an excellent overview of the statistical issues involved in combining effect size statistics and consider how to combine several metrics, including correlations (with and without the Fisher *z* transformation), standardized mean differences, differences between proportions, odds ratios and log odds ratios. Rather than review these methods, which are well known and widely used, this supplement will focus on a single metric, with the aim of illustrating the statistical concepts underlying meta-analysis. This metric is the simple mean difference (also called raw or unstandardized mean differences). As well illustrating the basics of meta-analysis, simple mean differences are among the most common effect sizes reported in published work. This type of meta-analysis has also been somewhat neglected in the literature. This neglect means that software for meta-analysis of simple mean differences is not widely available. Fortunately, basic meta-analytic calculations are not particularly difficult

for any of the common metrics (and most can be performed by hand or by using standard spreadsheet software). It seems logical to illustrate hand calculations using a meta-analytic method for which software is not widely available.

Although the emphasis here is on statistical issues, there are also many other important issues to consider. These include how to select studies for inclusion in a meta-analysis and how to deal with differences in the quality of the studies. Field and Gillett (2010) give a clear introduction to these issues, while Cooper and Hedges (1994) is one of the most comprehensive resources available for both the statistical and non-statistical issues involved in meta-analysis (e.g., the selection of studies for inclusion).

OS1.1.1 The case for meta-analysis of simple mean differences

Many meta-analyses involve combing effect sizes from studies involving differences in means (e.g., from paired or independent designs). If these studies all measure the outcome in exactly the same way then the most appropriate form of meta-analysis is that of the simple mean (or raw) differences. This has several advantages over using standardized metrics. Standardized metrics can obscure the meaning of the original variable (Shadish and Haddock, 1994). In particular, anything that influences the sample standard deviations (e.g., reliability, range restriction, differences in samples) will also influence standardized difference in means. This, in turn, can distort the meta-analysis (Baguley, 2009; Bond *et al.*, 2003). One solution is to correct for artifacts that distort standardized effect size (Hunter and Schmidt, 2004). Often the meta-analysis of simple mean differences is a better approach – particularly if the treatment or grouping variable is measured with little or no error. This happens to be the case for many meta-analyses involving experiments and quasi-experiments. The studies must also report a common measure, or measures, that can be converted to a common scale without standardization (e.g., using POMP scoring; Cohen *et al.*, 1999).

Shadish and Haddock (1994) describe methods for *fixed effect* and *random effects* meta-analysis assuming known variances. Such an approach works for standardized mean differences because these effects have $\sigma^2 = 1$. Bond *et al.* (2003) argue against the known variance approach (unless all studies have very large n). They demonstrate that the known variance approach can substantially underestimate the between-study variability of the effect. Methods proposed by Hartung and Knapp (2001) and Bond *et al.* (2003) avoid these difficulties, and illustrate many of the steps common to meta-analysis for standardized or simple effect size.

Most meta-analyses involve extensive work tracking down published and unpublished studies, deciding on inclusion criteria and calculating effect size estimates for each study. But sometimes a meta-analysis is more straightforward. For instance, a researcher in a specialist field may want to combine all the results from their own published and unpublished studies, or to combine effects in a multi-experiment paper (or from a PhD thesis). A major advantage of this approach is that it can reduce the risk of publication bias (by limiting the scope of the analysis to a known subset of studies in which no study has been excluded). The drawback is that this reduced scope limits the generality of the conclusions.

This approach involves five basic steps: i) collating relevant information about each study, ii) selecting the appropriate statistical model, iii) weighting the effect sizes, iv) estimating the variability of the effect and v) obtaining interval estimates or tests. For simple mean differences the relevant information about each effect includes summary statistics (e.g., sample size,

mean and variance of each sample) and contextual information to aid subsequent interpretation (e.g., key differences between the studies). The contextual information can be used to explore whether differences in study characteristics influence the size of an effect (known as a *moderator* analysis).

The choice of model is a fundamental decision in any meta-analysis and determines how the variability of the effect sizes is modeled (see Key Concept OS1). Weighting is a procedure that allows a researcher to adjust the estimate of the overall 'average' effect size according to the precision of each study (weighting more accurate estimates more heavily than less accurate ones). Estimating the variability of the effects is the heart of a statistical model in meta-analysis and makes it possible to obtain interval estimates and construct hypothesis tests. These inferences also require information about the distribution of the effects. Most meta-analytic methods assume samples from an independent, normal distribution. Further assumptions may be required for the specific method used in the analysis (e.g., the known variance assumption of the Shadish and Haddock equations).

KEY CONCEPT OS1

Fixed effect versus random effects analysis

The distinction between a fixed effect and random effects model is relevant to a number of statistical applications (see Hedges and Vevea, 1998; Shadish and Haddock, 1994). In a fixed effect model a single 'true' population parameter with a constant value is assumed (hence fixed *effect*, singular). When this population is sampled, each observation can be represented in terms of the population parameter θ ('theta') plus random error (usually assumed to be an independent, normal variable with $\mu = 0$ and unknown, constant variance σ^2):

$$y_j = \theta + \varepsilon_j \quad \varepsilon_j \sim N\left(0, \sigma^2\right) \qquad \text{Equation OS1.1}$$

If this looks familiar, it should. This is a very simple regression model (for a single parameter). The regression models considered thus far assume that the values of the predictors are fixed. In a basic regression model, only sampling error is treated as a random variable. For an independent measures design this means that the sampling units (e.g., participants) are treated as a random variable. If you measure each person only once, any individual differences are incorporated into the error term (and can't be separated from it).

In a random effects model, the population parameter itself varies (hence random *effects*, plural). The effect is no longer considered constant. Instead it is represented as a probability distribution (e.g., a normal distribution). Thus the random effects are represented by two parameters (the mean and the variance of the effects). This can be displayed as two separate equations. The first equation shows that the observed values are a function of θ and random error (as per the fixed effect model):

$$y_j = \theta_j + \varepsilon_j \quad \varepsilon_j \sim N\left(0, \sigma^2\right) \qquad \text{Equation OS1.2}$$

The second equation reveals that υ is also a random variable:

$$\theta_j = \theta_0 + \upsilon_j \quad \upsilon_j \sim N\left(0, \tau^2\right)$$

They can be combined into a single equation by substituting the second equation in place of θ_j:

$$Y_j = \theta_0 + \upsilon_j + \varepsilon_j \qquad \text{Equation OS1.3}$$

In this kind of random effects model there are two random variables (or two 'error' terms). One represents the sampling error ε_j and has variance σ^2. The other, υ_j, is the variability of the effect itself and has variance τ^2. This is a simple form of *multilevel model* (Hox, 2010; see also Chapter 18).

In meta-analysis, υ_j represents the variation in effect size between studies (not sampling error). This complicates the analysis, but is generally more plausible than a fixed effect model. In many meta-analyses it is unlikely that the studies are sufficiently similar that the effects are identical. Instead, they can be treated as a sample from a population of different effect sizes (sometimes termed a 'super-population' to distinguish it from the population sampled by each study). If the studies are very similar, a fixed effect model would still be a reasonable choice. However, the fixed effect model is limited in a very serious way – it does not generalize outside the studies in the analysis. With a fixed effect model there is no justification for inferences beyond the set of studies sampled, though there may be non-statistical grounds to generalize the findings.

When in doubt, employing a random effects model is considered a safer option. A fixed effect model assumes that the between-effect variance in the population is exactly zero. The random effects model includes an extra variance parameter and typically produces more conservative tests and wider interval estimates. Simulations suggest that it performs well even if the true population parameter is fixed (Hedges and Vevea, 1998).

It is possible to construct a significance test of the hypothesis that the effects in a meta-analysis are homogeneous (in which case a fixed effect model is appropriate) or heterogeneous (in which case a random effects model is appropriate). This is rarely appropriate. It is better to select the model based on *a priori* grounds, taking into account the context and aims of the research (Field and Gillett, 2010; Hedges and Vevea, 1998).

One guideline for distinguishing fixed from random effects is to consider the idea of a 'sampling fraction'. What proportion of the population of interest has been sampled? If the samples exhaust – or at least are a sizable fraction of – the population of interest (e.g., the studies include every type of situation you are interested in) then a fixed effect model may be appropriate. If the samples only represent a tiny fraction of the population of interest (e.g., if the studies only cover a few of the situations you are interested in) then the random effects model is preferable.

OS1.1.2 Fixed effect meta-analysis of simple mean differences

The first step in a basic meta-analysis of raw or simple mean differences is to obtain essential information for each effect. Assuming all the effects are computed between two independent means, this information is the size of each sample (n_1 and n_2), the difference in sample means ($\hat{\mu}_1 - \hat{\mu}_2$) and a standard error for each difference ($\hat{\sigma}_{\hat{\mu}_1 - \hat{\mu}_2}$). These are readily calculated from raw data or from summary data reported in most studies.

The population mean difference in the fixed effect model is estimated by the average individual 'study' effect plus random error with an independent, normal distribution. The estimate of the population mean difference ($\mu_1 - \mu_2$) can be denoted by θ and is usually computed as a weighted mean.[2] The fixed effect model described here is that proposed by Bond *et al.* (2003). The symbol G_j denotes the simple difference in means ($\hat{\mu}_{j1} - \hat{\mu}_{j2}$) for each of the $j = 1$ to J studies. The effect size estimate from a sample of J studies is:

$$G_j = \theta + \epsilon_j \quad \epsilon_j \sim N\left(0, \sigma^2\right) \qquad \text{Equation OS1.4}$$

If there are J studies with differences in means denoted as G_j, then W_j is the weight for j^{th} study and a weighted estimate of the population mean difference is:

$$\hat{\theta} = \frac{\sum_{j=1}^{J} \hat{W}_j G_j}{\sum_{j=1}^{J} \hat{W}_j} \qquad \text{Equation OS1.5}$$

Thus, each effect is multiplied by its weight, the weighted means are summed and then divided by the sum of the weights. If this weighting scheme is not intuitively obvious, think about what happens when all the weights are equal to one. In this case the numerator becomes the sum of all the differences and the denominator becomes J (the number of studies). Thus if all $W_j = 1$, the formula becomes the familiar arithmetic mean.

The same form of weighting can be applied to any form of effect size by replacing G_j with the statistic of interest. The weighting scheme can also be extended to include other factors. For instance, Shadish and Haddock (1994) provide a variant that also weights each study by a quality index: q. There are several methods for estimating q. A common approach is to list important features of a high-quality design and then set q equal to total number of features scored for each study. Although there are other ways to take into account the quality of a study (e.g., by excluding all low-quality studies) or treating quality as a moderator, weighting is one of the more flexible ones. To incorporate quality into the weighting you'd simply multiply each weight by the quality index for that study (i.e., replace $\hat{W}_j$ with $\hat{q}_j \hat{W}_j$ in both numerator and denominator of Equation OS1.5).

The formula for the weighted aggregate effect is common to many meta-analytic methods. A weighted estimate tends to be more accurate than an unweighted one if the effects are sampled at random from the super-population of interest. In extreme cases (e.g., where one study has a very large weight that overwhelms all others), an unweighted means analysis may outperform the weighted approach (Hunter and Schmidt, 2004). The correct weights to use depend primarily on whether a fixed effect or random effects model is used. In the fixed effect model the usual weight is the reciprocal of the variance of the sampling distribution of the effect (sometimes termed the *precision* of the effect). This variance is the square of the standard error of the effect size statistic (e.g., $\hat{\sigma}_{\hat{\mu}_{j,1}-\hat{\mu}_{j,2}} = \hat{\sigma}_{G_j}$ for a simple mean difference). The weights are therefore estimated from the standard error of the difference as:

$$\hat{W}_j = \frac{1}{\hat{\sigma}^2_{G_j}} \qquad \text{Equation OS1.6}$$

The crucial quantity $\hat{\sigma}_{G_j}$ is the denominator of the t statistic from an independent t test of the difference in means. This is frequently reported in published studies. Even if not reported, it can be readily calculated as G_j/t_j and so Equation OS1.6 can be rewritten as:

$$\hat{W}_j = \frac{1}{\left(G_j/t_j\right)^2} \qquad \text{Equation OS1.7}$$

Equation OS1.7 allows paired t tests and independent t tests to be combined in a single analysis. The method can also be used to carry out meta-analysis of one sample t tests (though it will not, as a rule, be reasonable to mix these with paired or independent mean differences). The main caveats are that the simple mean difference must be comparable between studies, and that the studies are independent of each other. The independence assumption would be

violated if, for instance, the same sample (e.g., a control group) were used to calculate more than one effect. If standardized effect size is used, then mixing paired and independent designs is rather awkward. This is because the standardizer is involved in both the calculation of the effect size statistic and the weights. Worse still, the standardizer for the weights must be different from that for the effects in order to make standardized paired differences comparable with independent differences (Hunter and Schmidt, 2004; Morris and DeShon, 2002).

Once a weighted mean has been computed, it is necessary to estimate its variance. Without knowing this, it will be difficult to interpret the weighted mean and it would not be possible to perform inference for the overall, aggregate effect. As the weights defined by Equation OS1.6 and Equation OS1.7 are formed from the sample variances, the weights can be used to estimate the variance of θ:

$$\hat{\sigma}^2_{\hat{\theta}} = \frac{1}{\sum_{j=1}^{J} \hat{W}_j} \left\{ 1 + \frac{4}{\left(\sum_{j=1}^{J} \hat{W}_j \right)^2} \times \sum_{L=1}^{J} \frac{(J-1)\hat{W}_L \sum_{j \neq L}^{J} \hat{W}_j}{(J-1)(\nu_L) - 4(J-2)} \right\} \qquad \text{Equation OS1.8}$$

The bracketed term on the right is not required for the known variance solution described by Shadish and Haddock (1994). Bond *et al*. (2003) show that when the population variance is estimated from the sample this additional term contributes a potentially large 'upward adjustment' to the estimate of $\hat{\sigma}^2_{\hat{\theta}}$. The term outside the bracket is simply the reciprocal of the sum of all J weights. Within the brackets the final term of the equation requires further explanation. Rather than indexing the studies and weights using j, a new index L is introduced. Like j, the index L is a set of numbers acting as labels for each of the J studies. The reason for the change in index is that the equation involves a nested calculation. Looking at the top half of the right-hand fraction you should see that it involves summing over all J studies except those where $L = j$. The nested calculation $\hat{W}_L \sum_{j \neq L}^{J} \hat{W}_j$ allows each study weight to be multiplied by the sum of all other study weights excluding itself. Without distinct index terms it would be hard to denote this operation.

Tests and interval estimates for $\hat{\theta}$ are constructed with the t distribution. The chief hurdle is determining the *df* (ν) for the fixed effect model. Bond *et al*. (2003) present the following formula, where ν_j indicates the *df* of the j^{th} study:

$$\nu_{\hat{\theta}} = \frac{\left(\sum_{j=1}^{J} \hat{W}_j \right)^2}{\sum_{j=1}^{J} \frac{\hat{W}_j^2}{\nu_j}} \qquad \text{Equation OS1.9}$$

The standard error of the fixed effect estimate is $\hat{\sigma}_{\hat{\theta}} = \sqrt{\hat{\sigma}^2_{\hat{\theta}}}$ and so a test statistic of the null hypothesis H_0: $\theta = 0$ is

$$t = \frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}} \sim t\left(\nu_{\hat{\theta}}\right) \qquad \text{Equation OS1.10}$$

and a CI for the fixed effect is:

$$\hat{\theta} \pm t_{v_{\hat{\theta}},1-\alpha/2} \times \hat{\sigma}_{\hat{\theta}} \qquad \text{Equation OS1.11}$$

Bond *et al.* (2003) argue that these estimates will be satisfactory provided $v_j \geq 8$ for each of the J studies.

Example OS1.1 Baguley *et al.* (2006, Experiment 1) report a meta-analysis of standardized mean differences comparing location memory between single anchor (SA) and paired single anchor (PSA) conditions. Subsequent examples will refer to this data set as the fixed effect meta-analysis data. In the SA conditions, participants had one opportunity to learn the location of an object, while in the PSA conditions participants had a second opportunity to learn the location of each object (relative to a different anchor or reference point). Interest focuses on whether performance in the PSA conditions is superior to that in the SA conditions. There are several possible choices of dependent variable for the study, but this example uses a measure termed *T* that ranges from zero to one (see Baguley *et al.*, 2006) and characterizes the proportion of information acquired about the object's location. To simplify hand calculation this example pools some data and reduces the number effects ('studies') from the first experiment to three. Summary data for these three studies are reported in Table OS1.1. Studies 1 and 2 involve incidental learning instructions while study 3 involves intentional learning (only in the latter are participants aware in advance that there is a memory test). Studies 2 and 3 use color stimuli while study 1 uses black and white stimuli.

Table OS1.1 Summary data for three independent group comparisons of SA and PSA conditions, for data adapted from Baguley *et al.* (2006)

	SA condition			PSA condition			
j	$\hat{\mu}$	$\hat{\sigma}$	***n***	$\hat{\mu}$	$\hat{\sigma}$	***n***	***t***
1	0.133	0.274	60	0.203	0.290	30	1.122
2	0.171	0.235	60	0.153	0.312	60	0.359
3	0.478	0.285	60	0.509	0.296	30	0.474

The first step is to obtain G_j and W_j for each of the $J = 3$ studies. As the summary data include t, Equation OS1.7 can then be used to derive $\hat{\sigma}_{G_j}$, $\hat{W}_j$, $G_j\hat{W}_j$ and $\hat{W}_j^2/v_j$. These values are reported in Table OS1.2.

Table OS1.2 Effect size estimates and intermediate results for the studies in Table OS1.1

j	G_j	v_j	$\hat{\sigma}_{G_j}$	$\hat{W}_j$	$G_j\hat{W}_j$	$\hat{W}_j^2/v_j$
1	0.070	88	0.0624	256.8	17.98	749.4
2	−0.018	118	0.0501	398.4	−7.17	1345.1
3	0.031	88	0.0654	233.8	7.25	621.2
S				889.0	18.06	2715.7

Getting the estimate of the fixed population effect involves plugging the effect sizes and weights into Equation OS1.5:

$$\hat{\theta} = \frac{\sum_{j=1}^{J} \hat{W}_j G_j}{\sum_{j=1}^{J} \hat{W}_j} = \frac{18.06}{889} \approx 0.020$$

This estimate illustrates the advantages of weighting. Study 2 has a larger sample and greater precision. It therefore influences $\hat{\theta}$ most heavily. An equally weighted estimate would be slightly larger. The weighted estimate of the difference is close to zero (given that the maximum range of a difference in proportion is −1 to 1).

The variance of the fixed effect estimate $\hat{\sigma}^2_{\hat{\theta}}$ is harder to calculate. Table OS1.3 sets out interim values for this calculation.

Table OS1.3 Interim values required to estimate the variance of the fixed effect estimate for the studies in Table OS1.1

j	$\sum_{j \neq L}^{J} \hat{W}_j$	$(J-1)\hat{W}_L \sum_{j \neq L}^{J} \hat{W}_j$	$(J-1)(\nu_L) - 4(J-2)$	$\frac{(J-1)\hat{W}_L \sum_{j \neq L}^{J} \hat{W}_j}{(J-1)(\nu_L) - 4(J-2)}$
1	632.2	324697.9	172	1887.8
2	490.6	390910.1	232	1685.0
3	655.2	306371.5	172	1781.2
Σ				5354.0

From Table OS1.2 the summed weights $\sum \hat{W}_j$ are 889 and (from Table OS1.3) the right-hand term sums to 5354. Plugging these values Equation OS1.8 produces

$$\hat{\sigma}^2_{\hat{\theta}} = \frac{1}{889} \times \left\{1 + \frac{4}{(889)^2} \times 5354\right\} \approx 0.00116$$

and (its square root) $\hat{\sigma}_{\hat{\theta}} \approx 0.034$. The *df* calculation requires the term $\sum \hat{W}_j / \nu_j$, which Table OS1.2 reports as 2715.7. Together, these produce:

$$\nu_{\hat{\theta}} = \frac{\left(\sum_{j=1}^{J} \hat{W}_j\right)^2}{\sum_{j=1}^{J} \frac{\hat{W}_j^2}{\nu_j}} = \frac{(889)^2}{2715.7} \approx 291.0$$

As $t_{291,.975}$ is approximately 1.968, the 95% CI for the fixed effect is $0.02 \pm .067$ or:

$$\hat{\theta} = .020, 95\% \text{ CI } [-.047, .087]$$

This result suggests an effect that is relatively small (compared to the full range of possible values) and includes zero as a plausible estimate. Baguley *et al.* (2006) consider the predictions of several theoretical models. One, an independence model (that assumes participants have access to two memories and attempt to draw on the second memory if retrieval of the first fails), predicts an effect of around .16. These data therefore exclude the independence model as a plausible explanation.

Pooling these studies into a single analysis assumes it is reasonable to lump them together. As all the studies had near identical experimental designs and similar samples this seems justified. In a strict sense, given that there are many potential variants of the experiment that have not been sampled, the fixed effect model limits generalization to the $J = 3$ effects in the analysis. This may not matter if the only goal is to combine experimental evidence from independent effects into a single aggregate estimate. Pooling similar studies in this way provides more precise estimates of the population effect size (and hence has greater statistical power).

OS1.1.3 Heterogeneity of effects

If a researcher is uncertain whether to apply a fixed effect or random effects model it is common to employ a test of homogeneity of effects (Shadish and Haddock, 1994). However, it is better to determine the choice of model on theoretical grounds rather than on the basis of a significance test (see Key Concept OS1). A homogeneity test is a test of the null hypothesis that all the effects are sampled from a population with the same fixed effect size (and, by implication, that the observed variability is sampling error). For standardized effect size metrics such tests usually take the form of a ratio of the deviations from the weighted mean effect size to the weighted variance estimate (producing a test statistic Q with an approximate χ^2 distribution). The ratio is small when the effect sizes are homogeneous (similar) and large when the effect sizes are heterogeneous (dissimilar). Thus statistical significance implies heterogeneity of effects. An obvious problem is lack of statistical power when J is small. For this reason it is probably safer to adopt a random effects model rather than rely on the outcome of a homogeneity test if the choice of model is not clear *a priori*.

The Q test (which assumes the population variance is known) is inappropriate for meta-analysis of simple mean differences. Bond *et al.* (2003) suggest using a statistic with an approximate F distribution (first proposed by Welch and termed F_{W}). Like the Q statistic, F_{W} is based on deviations from the weighted average effect:

$$F_W = \frac{(J+1)\sum_{j=1}^{J}\hat{W}_j\left(G_j-\hat{\theta}\right)^2}{J^2-1+2(J-2)u} \qquad \text{Equation OS1.12}$$

Here $\left(G_j-\hat{\theta}\right)^2$ is the squared deviation from the weighted mean effect for study j and u is:

$$u = \sum_{j=1}^{J}\frac{1}{v_j}\left(1-\frac{\hat{W}_j}{\sum_{L=1}^{J}\hat{W}_L}\right)^2 \qquad \text{Equation OS1.13}$$

Again this formula involves one sum (indexed by L) nested within another (indexed by j). This computation is not quite as complex, because the $\sum \hat{W}_L$ term is the sum of all J weights (with no effects excluded). An F statistic is a ratio of two χ^2 distributions, and has separate df for the numerator and denominator. For F_W these are $J - 1$ and $(J^2 - 1)/3u$ respectively.

An alternative approach to assessing heterogeneity is the analysis of moderator effects. This involves identifying potential predictors of variability in the population effect; attempting to model the variation explicitly rather than treating it as random variation (see Bonett, 2009). Bond *et al*. (2003) use the t distribution to construct CIs or tests of moderator effects by forming contrasts of effect sizes (see Section 15.6.1).

Example OS1.2 A test of homogeneity for the data in Table OS1.1 is made easier by calculating interim values required to determine u and F_W. These are set out in Table OS1.4.

Table OS1.4 Interim values to calculate F_w for the data in Table OS1.1

j	$\frac{\hat{W}_j}{\sum_{L=1}^{J} \hat{W}_L}$	$\frac{1}{v_j}\left(1 - \frac{\hat{W}_j}{\sum_{L=1}^{J} \hat{W}_L}\right)^2$	$G_j - \hat{\theta}$	$\hat{W}_j(G_j - \hat{\theta})^2$
1	0.2889	0.005746	0.050	0.642
2	0.4481	0.002581	−0.038	0.575
3	0.2630	0.006172	0.011	0.028
Σ		0.014499		1.245

As $u = 0.014499$ and $\sum \hat{W}_j \left(G_j - \hat{\theta}\right)^2 = 1.245$ the test statistic is:

$$F_W = \frac{(J+1)\sum_{j=1}^{J} \hat{W}_j \left(G_j - \hat{\theta}\right)^2}{J^2 - 1 + 2(J-2)u} = \frac{4 \times 1.245}{8 + 2 \times 0.014499} \approx 0.62$$

Given that $F < 1$ the test of heterogeneity is non-significant (as it implies that the between-effect variance is slightly less than expected under the null hypothesis of homogeneity). The df are $J - 1 = 2$ and $(J^2 - 1)/3u = 184.9$. Thus the test could be reported as: $F_W(2,184.9) = 0.62$, $p = .54$.

How should F_W be interpreted? There is little sign of heterogeneity, but with only three studies the test lacks statistical power. The argument for employing a fixed effect model should not rely on this test alone.

OS1.1.4 Random effects meta-analysis of simple mean differences

A random effects model assumes that the population effect size varies between studies rather than being fixed. Thus, in addition to within-study sampling error, it is necessary to estimate a second source of random variation. This is the between-study variation due to differences in the population effect size (see Key Concept OS1). For simple mean differences Bond *et al*.

(2003) adopt a random effects model proposed by Hartung and Knapp (2001). This estimates the between-study variance as:

$$\hat{\tau}^2 = \hat{\sigma}_G^2 - \frac{\sum_{j=1}^{J} \hat{\sigma}_{G_j}^2}{J} \qquad \text{Equation OS1.14}$$

Here $\hat{\sigma}_G^2$ is the unbiased estimate of the variance of the observed effects and $\hat{\sigma}_{G_j}^2$ is the sampling variance of the j^{th} effect. Hence $\hat{\tau}^2$ is the variance between effects not attributable to sampling error.

If the studies are a random sample of an infinite population of different effect sizes it is possible to estimate the weighted mean of this distribution as:

$$\tilde{\theta} = \frac{\sum_{j=1}^{J} \tilde{W}_j G_j}{\sum_{j=1}^{J} \tilde{W}_j} \qquad \text{Equation OS1.15}$$

This formula is identical to that for $\hat{\theta}$ except that the weights – now denoted by $\tilde{W}_j$ – incorporate between-effect variation:

$$\tilde{W}_j = \frac{1}{\hat{\sigma}_{G_j}^2 + \hat{\tau}^2} = \frac{1}{\left(G_j / t_j\right)^2 + \hat{\tau}^2} \qquad \text{Equation OS1.16}$$

The sampling variance of $\tilde{\theta}$ is estimated as:

$$\hat{\sigma}_{\tilde{\theta}}^2 = \frac{\sum_{j=1}^{J} \tilde{W}_j (G_j - \tilde{\theta})^2}{(J-1)\sum_{j=1}^{J} \tilde{W}_j} \qquad \text{Equation OS1.17}$$

The square root of this quantity is the standard error $\hat{\sigma}_{\tilde{\theta}}$. The ratio of $\tilde{\theta}$ to $\hat{\sigma}_{\tilde{\theta}}$ has a t distribution with $J - 1$ *df*. This permits to the construction of significance tests and interval estimates for the random effects model analogous to those in Equation OS1.10 and Equation OS1.11.

OS1.1.5 Selecting a meta-analytic model

Random effects models will, as a rule, produce wider CIs and more conservative tests than the fixed effect model (Hartung and Knapp, 2001; Bond *et al.*, 2003; Bonett, 2009). Despite this, the random effects approach is usually preferred over the fixed effect model because it is more robust to heterogeneity of effects (Hartung and Knapp, 2001; Field, 2005).

The random effects model does have its critics. Bonett (2008, 2009) has argued that the random effects model is inappropriate if effects cannot be considered a random sample from a super-population of effect sizes. He proposes an alternative fixed effect estimate of the

unstandardized or standardized mean differences of the observed studies based on unweighted means (Bonett, 2009). It employs a Welch-Satterthwaite correction and is robust with respect to unequal variances within samples (relevant in the independent groups case). Simulations suggest that this approach is also robust against heterogeneity of effect sizes. However, because a fixed effect estimate is adopted, the parameter estimated is not the mean of a super-population of effect sizes, but the unweighted mean of the set of studies included in the meta-analysis. An attractive feature of Bonett's approach is that heterogeneity of effects can be also be explored via regression methods.

The fixed effect model is recommended when aggregating effects from a small number of studies with very similar characteristics (e.g., from a multi-experiment paper). For studies with dissimilar characteristics a random effects model is an option, provided they can be considered a random sample from some super-population. If the studies are dissimilar, but not a random sample, then Bonett's robust fixed-effects approach should be explored.

OS1.2 Detecting potential problems in meta-analysis

Graphical methods can also be revealing about potential problems in meta-analysis, as well as being useful for communicating the results of the meta-analysis. Bax *et al*. (2009) review graphical tools and consider their effectiveness for detecting heterogeneity of effects and publication bias using a range of metrics. With respect to meta-analysis of simple mean differences, *decomposition plots* (Bond *et al*., 2003) can be used to explore the suitability of simple mean difference and standardized mean difference metrics.

OS1.2.1 Decomposition plots

In its basic form, the decomposition plot is a scatter plot with mean differences on the y-axis and $\hat{\sigma}_{pooled}$ on the x-axis. Lines corresponding to the average difference between studies or the average standardized difference can then be added. The average difference is depicted by a horizontal line. Plotting the average standardized difference is trickier. Bond *et al*. (2003) point out that:

$$\mu_1 - \mu_2 = \delta\sigma_{pooled} \qquad \text{Equation OS1.18}$$

As $\mu_1 - \mu$ is on the y-axis and σ_{pooled} on the x-axis this implies that the line $Y = 0 + \delta X$ can be added to represent the standardized mean difference (δ) on the decomposition plot.

Figure OS1.1 shows a decomposition plot for a meta-analysis of the effect of psychotherapy on length of hospitalization in days (Shadish and Haddock, 1994; Bond *et al*., 2003). Also plotted is the average simple mean difference between conditions (the dashed line) and the Hedges' g (the dotted line) as an estimate of δ. Unweighted averages are plotted (though weighted estimates can be plotted if preferred). An important feature of the plot is that both x-axis and y-axis use the same units (days of hospitalization). The scales of both axes must be constrained to be equal (to avoid introducing arbitrary distortions).

Bond *et al*. argued that the greater variability of points on the horizontal axis of Figure OS1.1 suggests that the simple (rather than standardized) mean difference is the best metric for these

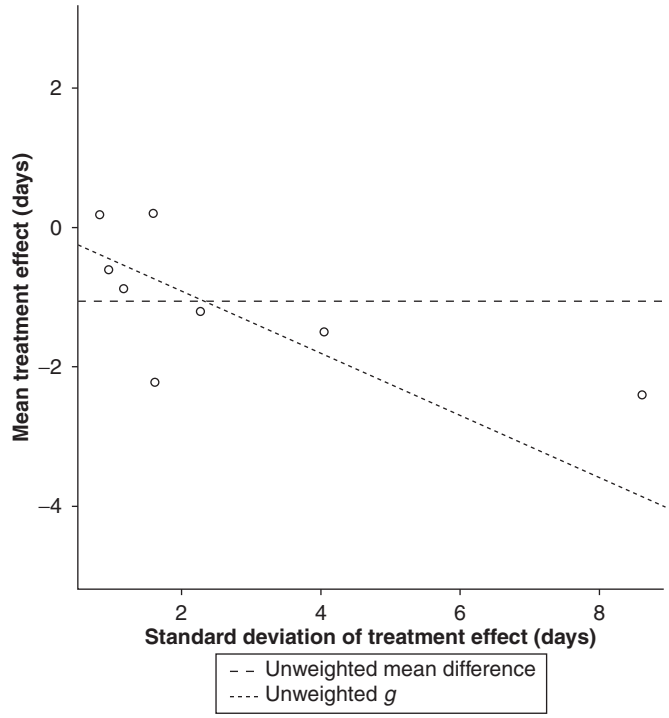

Figure OS1.1 Decomposition plot for the Shadish and Haddock (1994) psychotherapy data

data. The horizontal variability implies a noisy estimate of σ. Several $\hat{\sigma}$ values are also quite small, and these exert high leverage on the line $Y = 0 + gX$. Estimates of g will be extremely sensitive to low $\hat{\sigma}$ values. In addition, very small values of $\hat{\sigma}$ are likely to be underestimates of σ, because the sampling distribution of σ^2 is highly skewed (Browne, 1995; Vickers, 2003). The decision to use simple mean or standardized mean differences should not, however, rest simply on the decomposition plot. The decomposition plot may alert you to patterns among the sample estimates of σ that could distort g, but the decision boils down to two questions. What is the measure of most interest (theoretical or practical) and what is causing the variability in $\hat{\sigma}$? For instance, if the variability is caused by factors that are peripheral to the research question (e.g., reliability of the measures) then it should be treated as a nuisance variable to be removed from the estimate of the size of effect. Meta-analysis of the simple difference in means is one way to do this.

OS1.2.2 Detecting publication bias

The best-known method for detecting publication bias is the *funnel plot* (e.g., see Egger *et al.*, 1997). The original function of a funnel plot was to detect publication bias in meta-analysis (though it does so only indirectly). The plot is constructed with the observed effect size of each study on the x-axis and a measure of the precision of the studies on the y-axis. The measure of

precision is not the same for all funnel plots, with n, $\hat{\sigma}_{ES}$, $1/\hat{\sigma}_{ES}$ or $1/\hat{\sigma}^2_{ES}$ being potential options (where $\hat{\sigma}_{ES}$ is the standard error of the effect size statistic used in the meta-analysis).

Effect size statistics will be scattered around the average effect size estimate (usually plotted as a vertical line). Precisely measured effects should cluster more tightly round this average estimate than less precise estimates. If no publication bias is present, the points should fall in a roughly symmetrical pattern around the average effect. Assuming precision varies between studies (which it almost invariably does), this should produce a characteristic funnel shape: wide at the base and narrow at the top. Effects that are statistically non-significant are known to be harder to publish. Some of these effects are likely to be missing from the meta-analytic sample and will leave gaps in the plot. These gaps tend to appear where effects are most imprecise, at the base of the funnel, but only on one side. This is the side corresponding to an effect in the 'wrong' direction (e.g., showing that a treatment is ineffective). Thus publication bias should reveal itself as asymmetry in the plot.

It is important to emphasize that funnel plots and enhanced funnel plots do not assess publication bias directly – they detect asymmetry. Asymmetry can arise from sources other than publication bias such as heterogeneity of effect size or because the effect really is larger in small studies (e.g., because a treatment is easier to administer in small samples). Figure OS1.2a shows a funnel plot of the psychotherapy meta-analysis. The effect size plotted here is the simple mean difference. Precision is plotted from high to low (thus placing more precise studies – those with small standard errors – near the top). The vertical line shows weighted average effect size. Also displayed are approximate 95% confidence bands for the effect, illustrating the expected funnel shape. In this analysis, the small number of studies makes asymmetry in the plot difficult to spot, but the pattern is consistent with publication bias.

Figure OS1.2b shows an enhanced funnel plot using the *trim and fill* method (Duval and Tweedie, 2000). This estimates the number of 'missing' studies due to publication bias and uses the funnel plot to impute plausible values for them. Although it is tempting to add in the 'missing' studies, it is better to use this technique as a form of sensitivity analysis. Adding in the

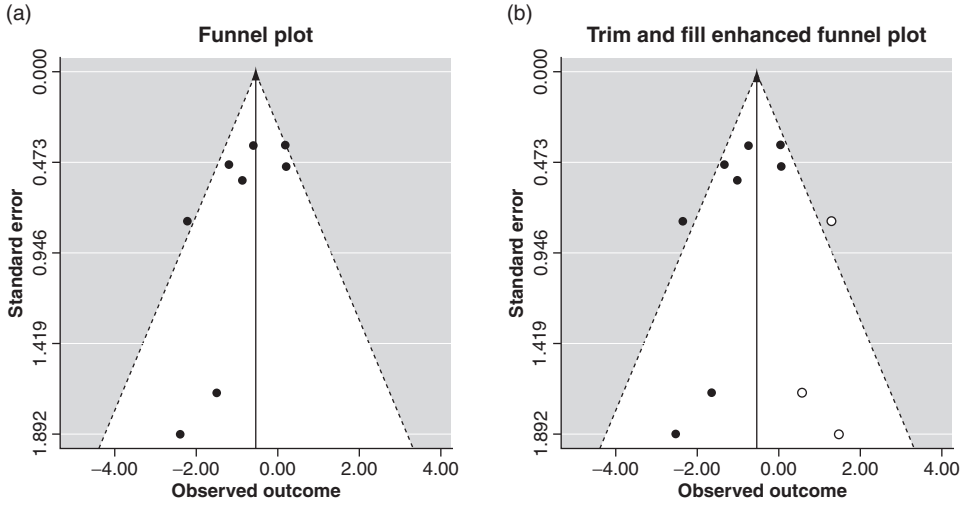

Figure OS1.2 Detecting asymmetry of effect size in the psychotherapy meta-analysis, using (a) a funnel plot with 95% confidence bands, or (b) a trim-and-fill enhanced funnel plot

studies might be appropriate if you were certain that publication bias was causing the asymmetry, but in most meta-analyses it could also be due to heterogeneity of effect size. A sensitivity analysis is preferred because it attempts to determine the impact on the conclusions of the meta-analysis if those non-significant studies were added to the data set (Peters *et al.*, 2008). Thus it is a way of checking the robustness of the analysis.

OS1.2.3 Heterogeneity of effect size

Bax *et al*. (2009) compare several methods for detecting heterogeneity of effect size, of which two will be described here. The first method is a normal probability plot of the residuals. This is not the best method for detecting heterogeneity, but has other uses (e.g., detecting potential outliers, spotting influential effects or assessing the assumption that the study effects are sampled from a normal distribution). The second method is a *forest plot*.

Forest plots are popular for reporting results. They depict the study effect sizes and the aggregate effect size as horizontal interval estimates with polygons for each point estimate. This produces a 'forest' of lines. A vertical line representing no effect can also be added. Thus the plot shows the variability within and between effects (as well as supporting inference about the overall average).

Figure OS1.3 shows a normal probability plot using the standardized residuals of a random effects model of simple mean differences (Hartung and Knapp, 2001; Viechtbauer, 2010). This plot suggests that the effects are within the expected range for the random effects model (all effects being within the confidence bands around the average effect). Note that the average

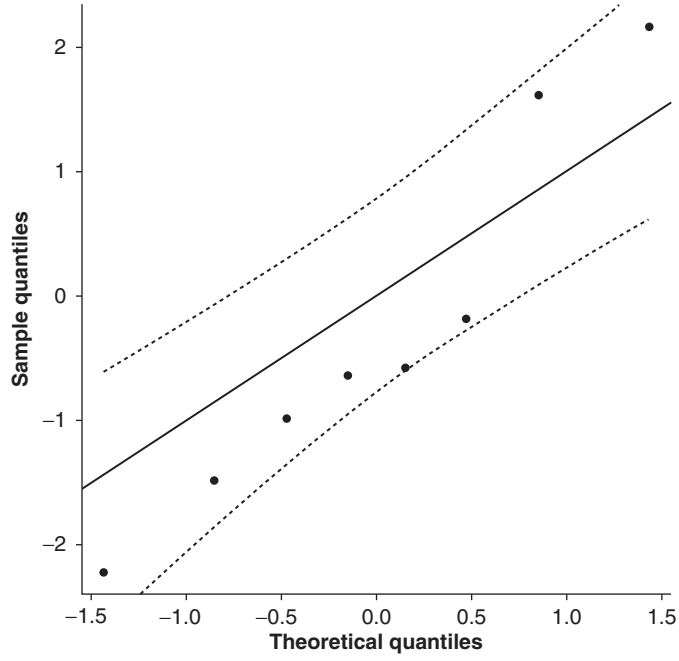

Figure OS1.3 Normal probability plot of standardized residuals, from a random effects meta-analysis of simple mean differences for the psychotherapy data

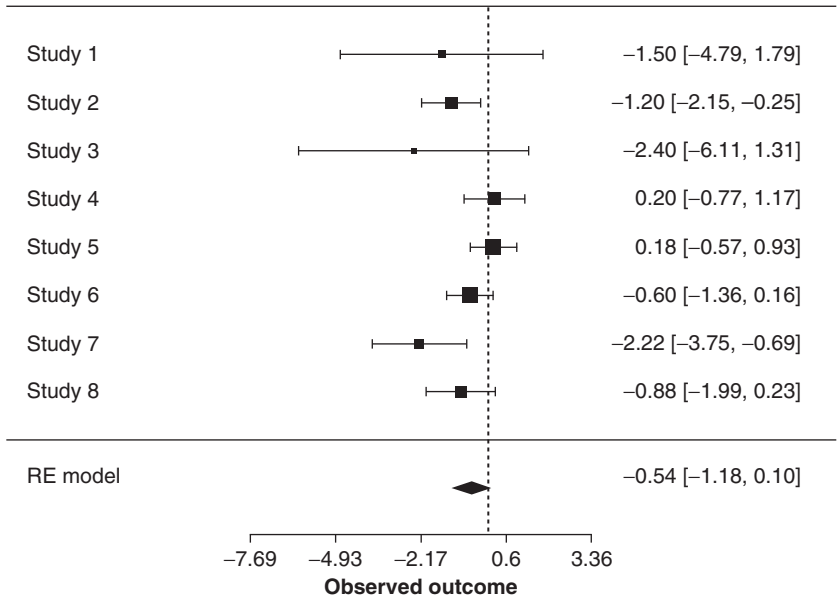

Figure OS1.4 Forest plot of the random effects meta-analysis of simple mean differences for the psychotherapy data

effect plotted here is a weighted average, and the residuals do not sum to zero. Several studies look to be quite influential, though this is almost inevitable with such a small set of studies. Figure OS1.4 shows a forest plot for the same analysis. This plot does not suggest heterogeneity of effects – all the effects lie within a fairly narrow range relative to the variability within studies. Of the individual effects, only studies 2 and 7 are statistically significant and the aggregate effect fails to reach statistical significance. Given the asymmetry in the funnel plot and the possibility of bias, the evidence of a treatment effect for these data is inconclusive.

OS1.2.4 Contrasts of simple mean differences in meta-analysis

In meta-analysis you may wish to explore not only an overall effect, but also the possibility that the effect is moderated by one or more other factors. Where the effects in question are differences among means and the moderator variable is categorical this allows a researcher to set up the hypothesis as a contrast. The contrast weights must, as usual, sum to zero and include at least one non-zero coefficient. For the meta-analysis of simple, raw mean differences Bond *et al*. (2003) proposed forming a test of the contrast with the formula:

$$t_{contrast} = \frac{\sum_{j=1}^{J} w_j G_j}{\sqrt{\sum_{j=1}^{J} \frac{w_j^2}{\hat{W}_j}}} \qquad \text{Equation OS1.19}$$

The numerator in Equation OS1.19 is the weighted mean effect size for the Bond *et al.* (2003) fixed effect model and J is the number of studies (i.e., mean differences in the meta-analysis). There is potential for confusion over the sets of weights denoted by w_j and $\hat{W}_j$. The w_j term refers to the contrast weights and $\hat{W}_j$ to the precision of the studies in the meta-analysis (i.e., the reciprocal of the sampling variance of the studies). The contrast has df determined as:

$$df_{contrast} = \frac{\left(\sum_{j=1}^{J} \frac{w_j^2}{\hat{W}_j}\right)^2}{\sum_{j=1}^{J} \frac{w_j^4}{\hat{W}_j^2 \nu_j}} \qquad \text{Equation OS1.20}$$

The squared correlation between effects and contrast weights can be used to give a sense of the proportion of between-study effect variance accounted for by the contrast. This is not a strict measure of the total variance (being unweighted by the precision of the effects), but is perhaps a fairer interpretation of the moderator effect's explanatory power than a weighted squared correlation (for the same reason $r^2_{alerting}$ calculated in this way is preferred to the version weighted by sample size in unbalanced designs).

Bonett (2009) favours moderator analysis as a way to explore between-study heterogeneity of effect size. His robust fixed effect method allows moderator analyses (including contrasts) to be defined as a general linear model. Baguley (2011) describes how to implement this model in R.

Example OS1.3 In the fixed effects meta-analysis reported in Example OS1.1, the first two studies used incidental recall and the third study used intentional recall. Is it possible that the effect size differs between intentional and incidental studies? This could be tested by a contrast using the weights $\{-0.5, -0.5, +1\}$. Using weights that have an absolute sum of two keeps the outcome scaled in terms of the original effect size metric. Combining these with the values from Table OS1.2, the t statistic is:

$$t_{contrast} = \frac{(-0.5 \times 0.07) + (-0.5 \times -0.018) + (1 \times 0.031)}{\sqrt{\left(\frac{0.25}{256.8} + \frac{0.25}{392.1} + \frac{1}{239.8}\right)}} = \frac{0.0047}{\sqrt{0.005781254}} = 0.06$$

This contrast is not statistically significant (because $t < 1$), but $df_{contrast}$ (required for the CI) are:

$$df_{contrast} = \frac{\left(\sum_{j=1}^{J} \frac{w_j^2}{\hat{W}_j}\right)^2}{\sum_{j=1}^{J} \frac{w_j^4}{\hat{W}_j^2 \nu_j}} = \frac{(0.005781254)^2}{0.000000211835} = 157.8.076$$

The df calculation is quite fiddly and hand calculation is best avoided.

The contrast reveals not even the barest hint of an effect. The CI for the contrast requires a critical value of t for two-sided $\alpha = .05$ and $df = 157.8$ (which is 1.975). The resulting 95% CI is [−0.15, 0.15]. This indicates at most a rather modest sized difference in means (given that the possible range of differences is −1 to 1), with a difference close to zero being highly plausible.

OS1.3 R code for Online Supplement 1

OS1.3.1 Meta-analysis of simple mean differences

Baguley (2011) provides R functions for the Bond *et al.* (2003) fixed effect method, the Hartung and Knapp (2001) random effects method and Bonett's (2009) method. The `metafor` package (Viechtbauer, 2010) also implements the Hartung and Knapp random effects method for independent group designs. The following commands run a random effects meta-analysis for the Shadish and Haddock length of hospitalization data (Shadish and Haddock, 1994, p. 273) using a vector of simple mean differences and a vector of standard errors as input:

```
m.e <- c(5, 4.9, 22.5, 12.5, 3.37, 4.9, 10.56, 6.5)
n.e <- c(13, 30, 35, 20, 10, 13, 9, 8)
sd.e <- c(4.7, 1.71, 3.44, 1.47, 0.92, 1.1, 1.13, 0.76)
m.c <- c(6.5, 6.1, 24.9, 12.3, 3.19, 5.5, 12.78, 7.38)
n.c <- c(13, 50, 35, 20, 10, 14, 9, 8)
sd.c <- c(3.8, 2.3, 10.65, 1.66, 0.79, 0.9, 2.05, 1.41)

diffs <- m.e - m.c
sd.pooled <- ((((n.c-1)*sd.c^2+(n.e-1)*sd.e^2)/(n.c+n.e-2)))^.5
se.diffs <- sd.pooled * sqrt(1/n.e + 1/n.c)

install.packages('metafor')
library(metafor)

rma.out <- rma(yi=diffs, sei=se.diffs, method='HE', knha=TRUE)
rma.out
```

This suggests that there is little heterogeneity in the population (the estimate of between-study variance is zero). The estimate of the effect is −0.54, 95% CI [−1.18, 0.10]. This is somewhat wider than the fixed effect meta-analysis of the same data reported by Bond *et al*. (2003). They report the estimate as −0.54, 95% CI [−0.95, −0.13]. The `metafor` package implements meta-analysis for a number of other metrics (and has excellent graphics output).

OS1.3.2 Diagnostic plots for meta-analysis

Figure OS1.1 provides a decomposition plot for the Shadish and Haddock (1994) psychotherapy data set, constructed as set out below. Baguley (2011) provides a function for a decomposition plot in R. They are fairly easy to construct. The first few commands below read in the raw data, extract simple mean differences and pooled standard deviations. The remaining code plots data and adds the required lines. An essential property of a decomposition plot is to match the scales of the two axes (because the simple mean differences and $\hat{\sigma}$ are on the same scale, the plot will be misleading if this relationship is not preserved). The graphics parameter `asp` sets the aspect ratio of the plot, and fixing this to one will keep the scales equal.

```
m.e <- c(5, 4.9, 22.5, 12.5, 3.37, 4.9, 10.56, 6.5)
n.e <- c(13, 30, 35, 20, 10, 13, 9, 8)
sd.e <- c(4.7, 1.71, 3.44, 1.47, 0.92, 1.1, 1.13, 0.76)
m.c <- c(6.5, 6.1, 24.9, 12.3, 3.19, 5.5, 12.78, 7.38)
n.c <- c(13, 50, 35, 20, 10, 14, 9, 8)
sd.c <- c(3.8, 2.3, 10.65, 1.66, 0.79, 0.9, 2.05, 1.41)

diffs <- m.e - m.c
sd.pooled <- ((((n.c-1)*sd.c^2+(n.e-1)*sd.e^2)/(n.c+n.e-2)))^0.5

plot(sd.pooled, diffs, xlab='Standard deviation of treatment
  effect (days)', ylab ='Mean treatment effect (days)', asp=1)

abline(mean(diffs), 0, lty=2)
abline(0, mean(diffs/sd.pooled), lty=3)

legend(4.5, 2.25, legend = c('Unweighted mean difference',
  expression(paste('Unweighted ', italic(g)))),
  lty=c(2,3),bty='n')
```

Several packages are available for meta-analysis, and a number provide funnel plots. Of these `metafor` (Viechtbauer, 2010) is among the most versatile, though for simple mean differences only the random effects model of Hartung and Knapp (2001) is recommended. The following commands create a funnel plot with 95% confidence bands for the Hartung and Knapp random effects model for simple mean differences from the psychotherapy data. The `funnel()` function uses a meta-analysis model object as input (which in turn requires the standard errors).

```
library(metafor)

se.diffs <- sd.pooled * sqrt(1/n.e + 1/n.c)
rma.out <- rma(yi=diffs, sei=se.diffs, method ='HE',
  knha=TRUE)

funnel(rma.out)
```

A trim and fill analysis can be obtained from the call `trimfill(rma.out)`, which produces a model object that can be used to create an enhanced funnel plot:

```
funnel(trimfill(rma.out))
```

The `metafor` package also includes a version of the `qqnorm()` function for meta-analysis model objects. Thus, once `metafor` is loaded, the following command produces a normal probability plot based on the standardized residuals of the meta-analysis:

```
qqnorm(rma.out)
```

The package also includes other plots, of which the forest plot is probably the most useful:

```
forest(rma.out)
```

Baguley (2011) describes how to use these functions to obtain forest plots and normal probability plots for the fixed effect and random effects meta-analysis methods proposed by Bond *et al*. (2003) and Bonett (2009).

OS1.3.3 Contrasts in meta-analysis

Enter the study data (including the weights for the effect size) from Example OS1.1 and create a vector of contrast weights:

```
Gj <- c(.07, -.018, .031)
Wj <- c(256.8, 392.1, 239.8)
nuj <- c(88, 118, 88)
wj <- c(-.5, -.5, 1)
```

The formulas for the *t* and the *df* are:

```
t.ma <- sum(wj*Gj)/sum(wj^2/Wj)^.5
df.contr <- sum(wj^2/Wj)^2/sum(wj^4/(Wj^2*nuj))
```

The 95% CI for the contrast is:

```
moe <- qt(.025,df.contr, lower.tail=FALSE) * sum(wj^2/Wj)^.5
ci.contrast <- c(sum(wj*Gj)-moe, sum(wj*Gj)+moe)
ci.contrast
```

Baguley (2011) includes functions that implement the Bond *et al*. fixed effect (2003) method (including contrasts).

R packages

Viechtbauer, W. (2010) Conducting Meta-analyses in R with the *metafor* Package. *Journal of Statistical Software*, 36, 1–48.

OS1.4 Notes on SPSS syntax for Online Supplement 1

OS1.4.1 Standardized effect size and meta-analysis

Calculating *d* family effect sizes will be easier by hand or using common spreadsheet software such as Excel than using SPSS syntax (point estimates for many *r* family metrics are easily obtained). SPSS syntax for obtaining CIs for standardized effect sizes including for δ is described by Smithson (2001) and Fidler and Thompson (2001). Field and Gillett (2010) describe macros for running meta-analysis of standardized effect size (*g* and *r*) using SPSS (with links to R for some output).

OS1.4.2 Diagnostic plots for meta-analysis

For details of meta-analytic procedures in SPSS it is worth looking at the SPSS macros in Field and Gillett (2010). Field and Gillett also explain how to obtain funnel plots (but not decomposition plots), and how to assess publication bias (using SPSS to invoke R).

OS1.5 Notes

1. The term 'study' is something of a misnomer. Meta-analysis combines effects that may come from different studies (but need not). However, it is the standard term in the literature to designate the source of an effect included in a meta-analysis and hence also adopted here.
2. Bond *et al*. use the symbol Δ. The symbol θ is adopted here to reduce confusion with other uses of Δ.

OS1.6 References

Baguley, T. (2009) Standardized or Simple Effect Size: What should be Reported? *British Journal of Psychology*, 100, 603–17.

Baguley, T. (2011) Meta-analysis of Simple Mean Differences: A Tutorial Review. Unpublished manuscript.

Baguley, T., Lansdale, M. W., Lines, L. K., and Parkin, J. (2006) Two Spatial Memories are not Better than One: Evidence of Exclusivity in Memory for Object Location. *Cognitive Psychology*, 52, 243–89.

Bax L., Yu, L. M., Ikeda, N., Fukui, N., Yaju, Y., Tsurata, H., and Moons, K. G. (2009) More than Numbers: The Power of Graphs in Meta-analysis. *American Journal of Epidemiology*, 169, 249–55.

Bond, C. F., Jr, Wiitala, W. L., and Richard, F. D. (2003) Meta-analysis of Raw Mean Differences. *Psychological Methods*, 8, 406–18.

Bonett, D. G. (2008) Meta-analytic Interval Estimation for Bivariate Correlations. *Psychological Methods, 13*, 173–189.

Bonett, D. G. (2009) Meta-analytic Interval Estimation for Standardized and Unstandardized Mean Differences. *Psychological Methods*, 14, 225–38.

Browne, R. H. (1995) On the Use of a Pilot Sample for Sample Size Determination. *Statistics in Medicine*, 14, 1933–40.

Cohen, P., Cohen, J., Aiken, L. S., and West, S. G. (1999) The Problem of Units and the Circumstance for POMP. *Multivariate Behavioral Research*, 34, 315–46.

Cooper, H., and Hedges, L. V. (1994) (eds) *The Handbook of Research Synthesis*. New York: Sage.

Duval, S. J., and Tweedie, R. L. (2000) A Nonparametric 'Trim and Fill' Method of Accounting for Publication Bias in Meta-analysis. *Journal of the American Statistical Association*, 95, 89–98.

Egger M., Smith, G. D., Schneider, M., and Minder. C. (1997) Bias in Meta-analysis Detected by a Simple, Graphical Test. *British Medical Journal*, 315, 629–34.

Fidler, F., and Thompson, B. (2001) Computing Correct Confidence Intervals for ANOVA Fixed- and Random-effects Effect Sizes. *Educational and Psychological Measurement*, 61, 575–604.

Field, A. P. (2005) Is the Meta-analysis of Correlation Coefficients Accurate when Population Effect Sizes Vary?. *Psychological Methods, 10,* 444–67.

Field, A. P., and Gillett, R. (2010) How to do a Meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63, 665–94.

Hartung, J., and Knapp, G. (2001) On Tests of the Overall Treatment Effect in Meta-analysis with Normally Distributed Responses. *Statistics in Medicine*, 20, 1771–82.
Hedges, L. V., and Vevea, J. L. (1998) Fixed- and Random-effects Models in Meta-analysis. *Psychological Methods*, 3, 486–504.
Hox, J. J. (2010) *Multilevel Analysis: Techniques and Applications* (2nd edn). New York/Hove: Routledge.
Hunter, J. E., and Schmidt, F. L. (2004) *Methods of Meta-analysis: Correcting Error and Bias in Research Findings* (2nd edn). Thousand Oaks, CA: Sage.
Morris, S. B., and DeShon, R. P. (2002) Combining Effect Size Estimates in Meta-analysis with Repeated Measures and Independent-groups Designs. *Psychological Methods*, 7, 105–25.
Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R. and Rushton, L. (2008) Contour-enhanced Meta-analysis Funnel Plots Help Distinguish Publication Bias from Other Causes of Asymmetry. *Journal of Clinical Epidemiology*, 61, 991–6.
Peters, J., Sutton, A. J., Jones, D. R., Abrams, K. R. and Rushton, L. (2010) Assessing Publication Bias in Meta-analysis in the Presence of Between-study Heterogeneity. *Journal of the Royal Statistical Society Series B*, 173, 575–91.
Shadish, W. R., and Haddock, C. K. (1994) Combining Estimates of Effect Size. In H. Cooper and L. V. Hedges (eds), *The Handbook of Research Synthesis*. New York: Sage, pp. 261–84.
Smithson, M. (2001) Correct Confidence Intervals for Various Regression Effect Sizes and Parameters: The Importance of Noncentral Distributions in Computing Intervals. *Educational and Psychological Measurement*, 61, 605–32.
Vickers, A. J. (2003) Underpowering in Randomized Trials Reporting a Sample Size Calculation. *Journal of Clinical Epidemiology*, 56, 717–20.
Viechtbauer, W. (2010) Conducting Meta-analyses in R with the *metafor* Package. *Journal of Statistical Software*, 36, 1–48.