A Bluffer's Guide to Meta-Analysis[1]

By

Dr. Andy Field

University of Sussex

What Is The Point of a Meta-Analysis?

Psychologists are typically interested in finding general answers to questions. For example, Lotze et al (2001) did a study to see what areas of the brain were activated during anal stimulation: they inserted balloons (not party ones) into people's rectums and inflated them while the person was in an fMRI scanner. Then they sang happy birthday and … OK, they didn't, but they really did do the balloon thing. One of the areas of the brain in which they were interested was the secondary somatosensory cortex (S2). Lotze et al. were probably interested in what brain regions were activated in their sample as a means of extrapolating to a wider population. However, what typically happens in science, is some other people then come along, they think 'hmm, shoving balloons up people's arses looks like a fun way to spend some research money' and off they go with their fMRI scanner and balloons to traumatise the local college populous. Of course, sooner or latter many more researchers will realise that this whole bum balloon thing is much more fun than whatever it is they're supposed to be doing, and before you know it, the literature is riddled with research papers (and the world is riddled with people who have conditioned surprised expressions on their face whenever they see an fMRI scanner). Can we assimilate all of these studies to improve the accuracy of our conclusions about which brain areas are activated by having crazy psychologists inflate balloons up our back passages?

Until about 30 years ago, the answer was simply to do a subjective evaluation of the literature. A typical review would entail the author collating articles on the given topic, summarising them and placing some kind of subjective weight on their findings. They might then, if you're lucky, conclude something about the topic of interest: perhaps that a certain area of the brain reliably lights up when your bottom is accosted by a balloon. These reviews have the obvious flaw that even the most discerning of researchers could give particular importance to studies that others might believe to be relatively less important. This can sometimes lead to quite long and heated debates in which different researchers reach different conclusions from the same literature. Meta-analysis rose out of a desire to objectify literature reviews using statistics. In short it is used to discover how big an effect actually is and what factors moderate that effect.

What Steps Do I have to Take?

When doing a meta-analysis you basically follow these steps:

Step 1: Do a Literature Search

The first step in meta-analysis is to search the literature for studies that have addressed the same research question (e.g. the ISI Web of Knowledge, PubMed, PsycInfo). We might also search relevant conference proceedings, hand-search relevant journals (in case the searches missed anything), search the reference sections of the articles that we have found, and consult

[1] Some of the material in this article were originally presented at a Psy-Pag organised one-day workshop on statistics at Oxford University, 15th April, 2005.

people we consider to be experts in the field – all of this is an attempt to avoid the file drawer problem (which we will discuss later on).

Step 2: Decide on some 'Objective' Criteria for Including Studies

OK, so we've got lots of studies, but obviously some of them might be useless. Badly conducted research can only serve to add bias into our meta-analysis, therefore, it's common to come up with some kind of inclusion criteria for studies. For example, in fMRI there are a variety of ways to process the enormous amounts of data that spew out, and you might reasonably decide that you'll include studies that follow a particular analysis protocol. Likewise, in a meta-analysis of a therapeutic intervention like cognitive behavioural therapy (CBT), you might decide on a working definition of what constitutes CBT, and maybe exclude studies that don't have proper control groups and so on. Your criteria will depend on what you're studying and any specific methodological issues in the field. You cannot exclude studies because you don't like the author. It is important that you formulate a precise set of criteria that is applied throughout, otherwise you may well be introducing subjective bias into the analysis.

It is also possible to classify studies into groups, for example methodologically strong or weak, and then see if this variable moderates the effect size (see Field, 2003a); by doing so you can see whether methodologically strong studies (by your criteria) differ in effect size to the weaker studies.

Step 3: Calculate the Effect Sizes

Once you have collected your articles, you need to find the effect sizes within them, or calculate them for yourself. I covered effect sizes (what they are, calculating them etc.) a few issues ago (see Field & Wright, 2006), so I won't re-explain them here.

Articles may not report effect sizes, or may report them in different metrics; your first job is to get effect sizes for each paper that represent the same effect and are expressed in the same way. If you were using *r* (my preferred effect size, and yes, you know you have officially become a dork when you have a favoured effect size measure), this would mean obtaining a value for *r* for each paper you want to include in the meta-analysis. A given paper may contain several *r*s depending on the sorts of questions you are trying to address with your meta-analysis. For example, I was recently involved in a meta-analysis of cognitive impairment in PTSD and 'cognitive impairment' was measured in a variety of ways in individual studies which meant I was often dealing with several effect sizes within a given article.

Step 4: Do the Meta-Analysis

This is the hard bit, which, if you've got to this stage, will seem ironic it'll probably have taken you most of your life to do steps 1 to 3. The main function of meta-analysis is to estimate the effect size in the population (the 'true' effect) by combining the effect sizes from a variety of articles. Specifically, the estimate is a weighted mean of the effect sizes. The 'weight' that is used is usually a value reflecting the sampling accuracy of the effect size. This makes statistical sense, because if an effect size has good sampling accuracy (i.e. it's likely to be an accurate reflection of reality) then it is weighted highly, whereas effect sizes that are a bit dodgy (are imprecise estimates) are given less weight in the calculations. Typically, as with any statistic, effect sizes based on large samples are more accurate reflections of the population than those based on small samples, the weight used is the sample size (or some function of it).

What can we get out of the meta-analysis?

- ✓ The 'true' effect size. That is the actual size of the effect in the population. For example, the true effect in the population of doing CBT on anxious children compared to waiting list controls. You can also compute confidence intervals for this true effect (wooppee!).
- ✓ The significance of the 'true' effect size. Actually, this isn't very interesting because significance is a function of sample size and so this really tells us nothing very useful

(see Field & Wright, 2006). Nevertheless, you can do it if you like (see Field, 2001 because I'm not going to explain it in this article).

- ✓ Meta-analysis can also be used to estimate the variability between effect sizes across studies (the homogeneity of effect sizes), but again, this in itself, isn't that interesting. There is accumulating evidence that effect sizes should be heterogenous across studies in the vast majority of cases (see, for example, the NRC paper, 1992). So, you can check if you like, but these tests of homogeneity typically have low power, and I'm of the view that unless there is evidence to the contrary, heterogenous effect sizes should be assumed.
- ✓ More interesting (no, really), is that given there is variability in effect sizes in most cases, this variability can be explored in terms of moderator variables (see Field, 2003a). For example, we might find that CBT including group therapy produces a larger effect size for improvement in eating disorders than CBT without a group component.

That's about it really.

Step 5: Write it up, lie back and Wait to see your first Psychological Bulletin Paper

Psychological Bulletin is one of the top ranking psychology journals in the universe. It is filled with meta-analyses. Meta-Analysis is the route to academic fame, fortune, the love of your department and the respect of your peers (or is that the other way around?)[2]. How do you write one up? Just follow Rosenthal's (1995) excellent guidelines; apart from being (as ever with Rosenthal) very sensible and very clearly-written, they were also published in Psychological Bulletin so they can hardly complain can they☺

How Do You Do A Meta-Analysis?

Ah, the tricky Step 4 eh? Well, obviously, there's just one way to do it, right? WRONG! This being statistics and everything there are numerous ways to do a meta-analysis, all of them are sort of different in different ways, involve making decisions about your data and have led some people (that'll be me then) to make small careers out of trying to establish which method is 'best'.

A Few of the More Important Issues to Bear in Mind

There are lots of issues to bear in mind and I've written about some of them (Field, 2001, 2003a, b; 2005a, b); to be fair, Schulze has written about them in more detail and rather more convincingly as have many others (Hunter & Schmidt, 2004; Rosenthal & DiMatteo, 2001). In terms of doing a meta-analysis, the main issues (as I see them) are:

1. Which Method Should I use?
2. Which conceptualisation of my data should I assume?

Actually, these two issues are linked. There are two ways to conceptualise meta-analysis: fixed effects and random effects models[3]. The fixed-effect model assumes that studies in the meta-analysis are sampled from a population in which the average effect size is fixed. Put another way, sample effect sizes should be homogenous because they come from the same population with a fixed average effect.

The alternative assumption is that the average effect size in the population varies randomly from study to study: studies in a meta-analysis come from populations that have different average effect sizes, so, population effect sizes can be thought of as being sampled from a

[2] At this point I should add that despite knowing this and despite having done lots of things involving meta-analysis, I've never actually done one and submitted it to Psychological Bulletin. Which just proves what an idiot I am.
[3] There are mixed models too, but I'm going to ignore them: see Overton, 1998.

'superpopulation'. See just about anything by me in the reading list for some further explanation. Put another way, the effect sizes should be heterogenous because they come from populations with varying average effect sizes.

How is this tied up with the method we use? Well, statistically speaking, the main difference between fixed- and random-effects models is in the amount of error. In fixed-effects models there is error introduced because of sampling studies from a population of studies. This error exists in random-effects models but there is *additional* error created by sampling the populations from a superpopulation (see Field, 2005b for some diagrams). So, calculating the error of the mean effect size in random-effects models involves estimating two error terms, whereas in fixed-effects models there is only one error term. This has some implications for computing the mean effect size.

The two most widely-used methods of meta-analysis are those by Hunter & Schmidt (2004) which is a random effects method, and the method by Hedges and Colleagues who provide both fixed- and random-effects methods.

I mentioned earlier on that there were rarely grounds to assume the fixed-effects case, that is, effect sizes are homogenous. You can trust me on this, or you can read the NRC (1992) report, or Hunter and Schmidt (2000) or Field (2005a) who argue or present data supporting this position. Despite overwhelming evidence that variable effect sizes are the norm in psychological data, this hasn't stopped lots of people from using fixed-effects methods. In fact, fixed effects methods are routinely applied to data even when effect sizes are variable (see Hunter & Schmidt, 2000) and this can have some fairly entertaining results such as a massive bias in resulting statistics (see Field, 2003b). To add to the confusion, the methods differ according to the effect size measure you use. I'm going to assume we're using *r*, but if you're using *d* you have to use slightly different equations (see Hedges & Vevea, 1999; Hunter & Schmidt, 2004).

Hedges and Colleagues' Method (Hedges & Olkin, 1985; Hedges & Vevea, 1998)

In this method, effect sizes are first converted into a standard normal metric (using Fisher's r-to-Z transformation) before calculating a weighted average of these transformed scores (in which *r* is the effect size from study *i*:

$$z_{r_i} = \tfrac{1}{2} Log_e\left(\frac{1+r_i}{1-r_i}\right) \qquad (1)$$

The transformation back to *r* is simply:

$$r_i = \frac{e^{(2z_i)}-1}{e^{(2z_i)}+1} \qquad (2)$$

In the fixed-effect model, the transformed effect sizes are used to calculate an average in which each effect size is weighted by the inverse within-study variance of the study from which it came (for correlation coefficients the sample size, *n*, minus three):

$$\bar{z}_r = \frac{\sum_{i=1}^{k} w_i z_{r_i}}{\sum_{i=1}^{k} w_i} = \frac{\sum_{i=1}^{k} (n_i - 3) z_{r_i}}{\sum_{i=1}^{k} (n_i - 3)} \ , \qquad (3)$$

in which *k* is the number of studies in the meta-analysis.

This average is used to calculate the homogeneity of effect sizes. The resulting statistic Q has a chi-square distribution with k – 1 degrees of freedom:

$$Q = \sum_{i=1}^{k} (w_i)(z_{r_i} - \bar{z}_r)^2 \qquad (4)$$

If you wanted to apply a fixed effects model you could stop here. However, as I've tried to convince you, this would be a bad thing, so read on. To calculate the random-effects average effect size, the weights use a variance component that incorporates both between-study variance and within-study variance. The between-study variance is denoted by τ^2 and is simply added to the within-study variance. The weighted average in the *z* metric is, therefore:

$$\bar{z}_r^* = \frac{\sum_{i=1}^{k} w_i^* z_{r_i}}{\sum_{i=1}^{k} w_i^*} \quad (5)$$

in which the weights $\left(w_i^*\right)$ are defined as:

$$w_i^* = \left(\frac{1}{w_i} + \tau^2\right)^{-1} \quad (6)$$

The between-study variance can be estimated in several ways (Hedges & Vevea, 1998; Overton, 1998), however, Hedges and Vevea use *Q* (which we came across earlier), *k*, and a constant, *c*:

$$\tau^2 = \frac{Q-(k-1)}{c} \quad (7)$$

where the constant, *c*, is defined (for correlation coefficients) as:

$$c = \sum_{i=1}^{k}\left(w_i\right) - \frac{\sum_{i=1}^{k}(w_i)^2}{\sum_{i=1}^{k}(w_i)} \ .$$

If τ^2 is negative then it is set to zero (because the variance between-studies cannot be negative). Having calculated τ^2, it is used to calculate the weights $\left(w_i^*\right)$, which in turn are used to calculate the mean effect size using equation 5. This average effect size must be converted back to the *r* metric (equation 2) before being reported.

Finally, it is useful to construct confidence intervals for the mean effect size (see Field, 2005c for a detailed explanation of confidence intervals and what they mean). To calculate these confidence intervals we need to know the standard error of the mean effect size is:

$$SE\left(\bar{z}_r^*\right) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}} \quad (8)$$

which uses the weights we've already calculated.

The confidence interval around the average effect size, is easily calculated using the standard error and the two-tailed critical value of the normal distribution (which is 1.96 for the most commonly used 95% confidence interval). The upper and lower bounds are calculated by taking the average effect size and adding or subtracting its standard error multiplied by 1.96:

$$CI_{Upper} = \bar{z}_r^* + 1.96SE\left(\bar{Z}_r^*\right) \quad (9)$$

$$CI_{Lower} = \bar{z}_r^* - 1.96SE\left(\bar{Z}_r^*\right) \quad (10)$$

These values are again transformed back to the *r* metric before being reported.

Hunter and Schmidt Method (Hunter & Schmidt, 2004)

Although this method's greatest virtue is its emphasis on isolating and correcting for sources of error such as sampling error and reliability of measurement variables, it is dealt with here in only its simplest form. Unlike Hedges' method the untransformed effect-size estimates, *r*, are

used to calculate the weighted mean effect size, and the weight used is simply the sample size, *n*:

$$\bar{r} = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i} \quad (11)$$

Hunter and Schmidt (2004) argue that the variance across sample effect sizes consists of the variance of effect sizes in the population and the sampling error and so the variance in population effect sizes is estimated by correcting the variance in sample effect sizes by the sampling error. The variance of sample effect sizes is the frequency weighted average squared error:

$$\sigma_r^2 = \frac{\sum_{i=1}^{k} n_i (r_i - \bar{r})^2}{\sum_{i=1}^{k} n_i}. \quad (12)$$

The sampling error variance is calculated as:

$$\sigma_e^2 = \frac{\left(1-\bar{r}^2\right)^2}{\bar{N}-1} \quad (13)$$

in which, is the average effect size, and is the average sample size. The variance in population effect sizes is estimated by subtracting the sampling error variance from the variance in sample effect sizes:

$$\hat{\sigma}_\rho^2 = \sigma_r^2 - \sigma_e^2 \quad (14)$$

Hunter and Schmidt recommend correcting this estimate for artefacts (see Hunter & Schmidt, 2004) and then constructing credibility intervals. These intervals are based on taking the average effect size and adding or subtracting from it the square root of the estimated population variance multiplied 1.96 (for a 95% interval):

$$\text{Credibility Interval}_{\text{Upper}} = \bar{r} + 1.96\sqrt{\hat{\sigma}_\rho^2} \quad (16)$$

$$\text{Credibility Interval}_{\text{Lower}} = \bar{r} - 1.96\sqrt{\hat{\sigma}_\rho^2} \quad (17)$$

An Example

In my last Bluffer's guide on effect sizes, I used an example of whether listening to Cradle of Filth (CoF) turns people into Granny-murdering devil-worshippers. In that example, we exposed unborn children to Cradle of Filth (or not) and observed how they turn out years later. Now clearly, this is a topic that would interest lots of researchers so lets imagine lots of researchers had addressed a similar question (perhaps using different methodologies, and different outcome measures). We can follow the steps outlined above:

Step 1: Do a Literature Search

Ok, we searched the ISI Web of Knowledge, PubMed, PsycInfo etc. and found the studies listed in Table 1.

Table 1: Summary of articles found on CoF and satanic activity.

Study	Journal	Measures	Rating/Comment
Incon & Tennent(2002)	Knitting Pattern Review	Grannies Murdered	***** (Nice Stats)
Little, Bo & Peep (2002)	Journal of Sacrificial Goats	Goats Sacrificed	***** (Nice Graphs)
Beelzibub (2003)	Norwegian Journal of Satanism	Daemons summoned	***** (Good use of GLMs)
Osbourne (2004)	Bird Eater Monthly	Doves eaten	***** (Helpful effect-sizes)
Fisher[4] (2004)	Goat Lover monthly	Goats abducted	****
Field & Hedgerow (2003)	Latvian Journal of Shite Research	Pentagrams drawn	* A bit iffy really …

Step 2: Decide on some 'Objective' Criteria for Including Studies

You might add some ratings based on your systematic criteria (see Table 1) to help you identify features of the study, or generally assess the quality of the research). Exclusions might also have to be made on the grounds that the authors do not report enough information for effect sizes to be computed. In this case we'll exclude the Field & Hedgerow article on the basis that everything I do is rubbish.

Step 3: Calculate the Effect Sizes

Next, we calculate the effect sizes for each study, and maybe tabulate them with other helpful information (such as the sample size on which the effect size is based, *N*) – Table 3.

Table 3: Effect sizes and sample sizes for remaining studies.

Study	Journal	Measures	*N*	Effect (*r*)
Incon & Tennent(2002)	Knitting Pattern Review	Grannies Murdered	135	-.68
Little, Bo & Peep (2002)	Journal of Sacrificial Goats	Goats Sacrificed	1235	-.79
Beelzibub (2003)	Norwegian Journal of Satanism	Daemons summoned	570	-.74
Osbourne (2004)	Bird Eater Monthly	Doves eaten	190	.12
Fisher (2004)	Goat Lover Monthly	Goats abducted	52	.24
Field & Hedgerow (2003) Excluded for Being Shite				

[4] Apparently, Fisher (yes, the one who invented lots of things like ANOVA) had a pet Goat that he took to work with him.

Step 4: Do the Meta-Analysis

Let's first, do the Hunter-Schmidt method because this is, frankly, easier.

We can first extend our table to compute the weighted effect sizes (Table 4):

Table 4: Weighted effect sizes for the studies in our meta-analysis.

Study	N	r	Nr
Incon & Tennent(2002)	135	-.68	-91.80
Little, Bo & Peep (2002)	1235	-.79	-975.65
Beelzibub (2003)	570	-.74	-421.80
Osbourne (2004)	190	.12	22.80
Fisher (2004)	52	.24	12.48
Total	**2182**		**-1452.97**

Using equation 11, we get a weighted mean of –.666:

$$\bar{r} = \frac{-1452.97}{2182}$$
$$= -.666$$

This means that the true effect in the population is a strong negative effect of listening to CoF on satanic behaviour (i.e. the more you listen to CoF, the less goat abducting, granny murdering etc. you engage in).

To calculate the credibility interval, we can use the true effect we've just calculated to create a new table (Table 5).

Table 5: Calculating Credibility Intervals

Study	N	r	$(r-(-.666))^2$	$n(r-(-.666))^2$	Nr
I & T (2002)	135	-.68	.000196	0.26	-91.80
L, B & P (2002)	1235	-.79	.015376	18.99	-975.65
B (2003)	570	-.74	.005476	3.12	-421.80
O (2004)	190	.12	.617796	117.38	22.80
F (2004)	52	.24	.820836	42.68	12.48
Total	2182			182.43	-1452.97

Column 4 (and its sum) represent the entire top half of equation 12 (column 3 is a stepping stone to get there), the sum of column 2 is the bottom half of this equation. This gives us (for equation 12):

$$\hat{\sigma}_r^2 = \frac{182.43}{2182} = 0.0836$$

The sampling error variance (equation 13) can be obtained more directly. The only value we don't have is the average sample size, but we have the total from Table 5 so we can just divide this by the number of studies (in this case 5). We get for equation 13:

$$\begin{aligned}\hat{\sigma}_e^2 &= \frac{\left(1-\bar{r}^2\right)^2}{\bar{N}-1} \\ &= \frac{\left(1-\left(-.666\right)^2\right)^2}{\frac{2182}{5}-1} \\ &= 0.000711\end{aligned}$$

The two values we've just calculated are then used to estimate the variance of population effect sizes (equation 14):

$$\begin{aligned}\hat{\sigma}_\rho^2 &= \hat{\sigma}_r^2 - \hat{\sigma}_e^2 \\ &= 0.0836 - 0.000711 \\ &= 0.0829\end{aligned}$$

This value is then used to create 95% credibility intervals (equations 16 and 17):

$$\text{Credibility Interval}_{\text{Upper}} = \bar{r} + 1.96\sqrt{\hat{\sigma}_\rho^2} = -.10$$

$$\text{Credibility Interval}_{\text{Lower}} = \bar{r} - 1.96\sqrt{\hat{\sigma}_\rho^2} = -1.23$$

Ok, that's the easy one out of the way, now Hedges' method. I've rounded off the decimal places, but obviously accuracy is important so when doing the actual calculations I have kept all decimal places in, so you might get slightly different values if you just use my rounded off values from the tables. However, you can download an Excel spreadsheet with formulae for this method. It uses these data, but you can easily extend it to your own meta-analysis by adding rows to the tables: you might have to tweak some of the formula so that they refer to the correct cells though: http://www.sussex.ac.uk/Users/andyf/pgstat.html)

Table 6 shows the main steps in the fixed-effects method. Again, we have our studies, our sample sizes (*n*) and our effect sizes (*r*). The first thing is to calculate the weights in equation 3. This is simple, just subtract 3 from the sample size (see column 3). Next, we have to convert *r* into the *z* metric using equation 1 (this is shown in column 6). Finally, we times these z values by the weights (i.e. we multiply column 3 with column 6) and this gives us column 7. The sum of column 7 is the top half of equation 3, and the sum of column 3 is the bottom half[5]. Therefore, to get the true effect size, we use these two sums:

$$\bar{z}_r = \frac{-1933.83}{2167} = -0.8924$$

Remember that this is in the z-metric, so we must convert back to *r* using equation 2, which gives us a true effect of –0.7126.

[5] You might be wondering what column 4 is all about (the weights squared), these values come in useful later on.

Table 6: Hedges' Fixed-effects meta-analysis

Study	*n*	*w*	w^2	*r*	z_r	$W \times z_r$
I & T (2002)	135	132	17424	-0.68	-0.83	-109.44
L, B & P (2002)	1235	1232	1517824	-0.79	-1.07	-1320.00
B (2003)	570	567	321489	-0.74	-0.95	-538.92
O (2004)	190	187	34969	0.12	0.12	22.55
F (2004)	52	49	2401	0.24	0.24	11.99
		2167	1894107			-1933.83

We need this value to calculate Q, the homogeneity of effect sizes in equation 4. Table 7 shows the steps here. We use the weights from each study again ($N - 3$), which is the left half of equation 4. The right hand side consists of the effect size (in z units) minus the mean effect size that we just calculated (−0.8924)—see column 4 of table 7. We then square these values (Column 5) and then multiply by the weights (Column 6). We sum the resulting values to give us the result of equation 4, $Q = 297.18$.

Table 7: Calculating the Homogeneity of Effect Sizes

Study	***w***	z_r	$(z_r)-(-0.892)$	**(Column 4)^2**	***W* × Column 5**
I & T (2002)	132	-0.83	0.06	0.004	0.53
L, B & P (2002)	1232	-1.07	-0.18	0.032	39.49
B (2003)	567	-0.95	-0.06	0.003	1.91
O (2004)	187	0.12	1.01	1.026	191.89
F (2004)	49	0.24	1.14	1.293	63.36
	2167				**297.18**

We now have to use this value of Q to calculate tau^2 (equation 7). However, first we have to calculate c, and we already have all of the values we need in Table 6; that is, the sum of weights (bottom of column 3) and the sum of weights squared (bottom of column 4). This gives us:

$$c = \sum_{i=1}^{k}(w_i) - \frac{\sum_{i=1}^{k}(w_i)^2}{\sum_{i=1}^{k}(w_i)}$$

$$= 2167 - \frac{1894107}{2167}$$

$$= 1292.93$$

Using Equation 7, we can get tau^2 as (*k* is the number of studies in the meta-analysis, and Q we have just computed):

$$\tau^2 = \frac{Q-(k-1)}{c} = \frac{297.18-(5-1)}{1292.93} = 0.227$$

We can now use this value to calculate the weights for the random-effects model (equation 6). Again, you can break this equation down into a table (Table 8).

Table 8: Calculating the random-effects mean effect size.

Study	*N*	w	1/w	*1/w + tau^2*	1/Column 5	z_r	(W*) × z_r
I & T (2002)	135	132	0.008	0.23	4.30	-0.83	-3.562
L, B & P (2002)	1235	1232	0.001	0.23	4.42	-1.07	-4.740
B (2003)	570	567	0.002	0.23	4.41	-0.95	-4.188
O (2004)	190	187	0.005	0.23	4.34	0.12	0.523
F (2004)	52	49	0.020	0.25	4.07	0.24	0.997
		2167			21.53		-10.97

Table 8 begins by breaking down equation 6. First we calculate the original weights (*n* – 3), we then divide 1 by these values (1/(*n*–3)), next we add our previously calculated value of tau^2 (0.227) to give us the inside of the brackets in equation 6. Finally, we divide 1 by these values and end up with the weights for the random-effects method (column 6). We already know the z-score values of the effect sizes, so we need to simply multiply these new values to give us the new weighted effect sizes (column 8). The mean is derived from equation 5, the top half of which is the sum of column 8 in Table 8 and the bottom half of which is the sum of column 6 in Table 8:

$$\bar{z}_r^* = \frac{\sum_{i=1}^{k} w_i^* z_{r_i}}{\sum_{i=1}^{k} w_i^*} = \frac{-10.97}{21.53} = -0.509$$

Remember that this is in the z-metric, so we must convert back to *r* using equation 2, which gives us a true effect of –0.469. Notice, this result is a fair bit smaller than the fixed-effects model!

To get the confidence intervals we use equation 8 to find the standard error of our mean effect size; the sum of weights is 21.53 (see Table 8), which gives us:

$$SE\left(\bar{z}_r^*\right) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}} = \sqrt{\frac{1}{21.53}} = 0.215$$

To get the 95% confidence intervals we times this by 1.96 and then either add or subtract it from the mean effect size:

$$CI_{Upper} = \bar{z}_r^* + 1.96SE\left(\bar{Z}_r^*\right) = -0.509 + 0.422 = -0.087$$

$$CI_{Lower} = \bar{z}_r^* - 1.96SE\left(\bar{Z}_r^*\right) = \text{-0.509-0.422 = -0.931}$$

Again, these values are in z units so we have to convert back to *r* using equation 2, which gives us an upper CI of -0.087 (it hasn't changed to 3 decimal places because it is close to zero) and a lower value of –0.731.

As with the Hunter Schmidt method, the results show that listening to CoF has a negative effect on satanic behaviours, but notice that the estimate of the true effect is lower than when the Hunter-Schmidt method was used.

Step 5: Write it up, lie back and Wait to see your first Psychological Bulletin Paper

Or spend the rest of your life in an asylum after all of that maths☺

That's All Bloody Hard Work, Can't I just get SPSS to Do It?

No actually (well, not unless you write your own syntax)! Hunter and Schmidt do provide software on the CD ROM in their book for doing their method. There are also (I believe) some commercially available packages that do it, but frankly they're a bit of a rip-off: it's not *that* hard.

Which Method is Best?

Our two analyses yielded different results which beg the question of which method is most accurate. Several people have compared methods: Schulze (2004) did a whole book of simulations that compared these and other methods and it's well worth a look. I have also published a couple of papers doing similar things (but less extensively than Schulze): Field, 2001; 2005a. We are not the only people to have studied these issues but given I've already written a very long Bluffer's guide, I'll just say, if you're interested, look at these sources for reviews of the evidence.

Other Problems with Meta-Analysis

Publication Bias and The 'File Drawer' Problem: Publication bias refers to the fact that significant findings are more likely to be published than non-significant findings. It is sometimes known as the 'file drawer' problem because non-significant research is more likely to end up in the researchers file drawer than in a journal. This bias can be substantial, estimates suggest that significant findings are eight times more likely to be submitted than non-significant ones and 97% of articles in psychology journals reported significant results. The effect of this bias is that meta-analytic reviews are likely to over-estimate mean effect sizes.

Artefacts: Effect sizes are influenced by the quality of the research and quality precision of measurement of variables. The error in the measurement of variables will vary across studies and correlational research is prone to variance in the range of scores elicited from participants (range variation). Hunter and Schmidt (1990) have suggested statistical techniques for correcting for measurement error and range variation and you are well-advised to consult their book.

Further Reading

The following are good introductions to the issues in meta-analysis (in alphabetic order): Field (2005b, 2003a), Hunter & Schmidt (2004); Rosenthal (1990), Rosenthal & DiMatteo (2001).

References

Field, A. P. (2005a). Is the meta-analysis of correlation coefficients accurate when population effect sizes vary? Psychological Methods, 10 (4), 444-467.

Field, A. P. (2005b). Meta-analysis. In J. Miles & P. Gilbert (eds.) A handbook of research methods in clinical and health psychology (pp. 295-308). Oxford: Oxford University Press.

Field, A. P. (2003a). Can meta-analysis be trusted? The Psychologist, 16, 642-645.

Field, A. P. (2003b). The problems in using Fixed-effects models of meta-analysis on real-world data. Understanding Statistics, 2, 77–96.

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. Psychological Methods, 6 (2), 161–180.

Field, A. P., & Wright, D. B. (2006). A bluffer's guide to effect sizes. Psy-Pag Quarterly, 58, 9-23.

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

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

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: implications for cumulative knowledge in Psychology. *International Journal of Selection and Assessment*, *8*, 275–292.

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

Lotze, M., Wietek, B., Birbaumer, Ehrhardt, J., Grodd, W., & Enck., P. (2001). Cerebral activation during anal and rectal stimulation. NeuroImage, 14, 1027-1034.

National Research Council (1992). Combining information: Statistical issues and opportunities for research. Washington, D.C.: National Academy Press.

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

Rosenthal, R. (1995). Writing meta-analytic reviews. Psychological Bulletin, 118, 183-192.

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

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods and literature reviews. Annual Review of Psychology, 52, 59-82.

Schulze, R. (2004). Meta-analysis: a comparison of approaches. Cambridge, MA: Hogrefe & Huber.