British Journal of Mathematical and Statistical Psychology (2010), 63, 665–694

The British Psychological Society
www.bpsjournals.co.uk

Expert tutorial

How to do a meta-analysis

Andy P. Field[1]* and Raphael Gillett[2]*
[1]School of Psychology, University of Sussex, UK
[2]School of Psychology, University of Leicester, UK

Meta-analysis is a statistical tool for estimating the mean and variance of underlying population effects from a collection of empirical studies addressing ostensibly the same research question. Meta-analysis has become an increasing popular and valuable tool in psychological research, and major review articles typically employ these methods. This article describes the process of conducting meta-analysis: selecting articles, developing inclusion criteria, calculating effect sizes, conducting the actual analysis (including information on how to do the analysis on popular computer packages such as IBM SPSS and R) and estimating the effects of publication bias. Guidance is also given on how to write up a meta-analysis.

What is meta-analysis and how do I do it?

Psychologists are typically interested in finding general answers to questions across this diverse discipline. Some examples are whether cognitive behaviour therapy (CBT) is efficacious for treating anxiety in children and adolescents (Cartwright-Hatton, Roberts, Chitsabesan, Fothergill, & Harrington, 2004), whether language affects theory of mind performance (Milligan, Astington, & Dack, 2007), whether eyewitnesses have biased memories of events (Douglass & Steblay, 2006), whether temperament differs across gender (Else-Quest, Hyde, Goldsmith, & Van Hulle, 2006), the neuropsychological effects of sports-related concussion (Belanger & Vanderploeg, 2005), and how pregnant women can be helped to quit smoking (Kelley, Bond, & Abraham, 2001). These examples illustrate the diversity of questions posed by psychologists to understand human behaviour. Although answers to these questions can be obtained in single pieces of research, when these studies are based on small samples the resulting estimates of effects will be more biased than in large-sample studies. Also, replication is an important means to deal with the problems created by

*Correspondence should be addressed to Professor Andy P. Field, School of Psychology, University of Sussex, Falmer, Brighton BN1 9QH, UK (e-mail: andyf@sussex.ac.uk); or Raphael Gillett, School of Psychology, Henry Wellcome Building, University of Leicester, Lancaster Road, Leicester LE1 9HN, UK (e-mail: rtg@le.ac.uk).

DOI:10.1348/000711010X502733

measurement error in research (Fisher, 1935). For these reasons, different researchers often address the same or similar research questions, making it possible to answer questions through assimilating data from a variety of sources using meta-analysis. A meta-analysis can tell us several things:

(1) *The mean and variance of underlying population effects*. For example, the effects in the population of doing CBT on anxious children compared to waiting-list controls. You can also compute confidence intervals for the population effects.
(2) *Variability in effects across studies*. Meta-analysis can also be used to estimate the variability between effect sizes across studies (the homogeneity of effect sizes). Some meta-analysts report these statistics as a justification for assuming a particular model for their analysis or to see whether there is variability in effect sizes that moderator variables could explain (see Step 4). However, there is accumulating evidence that effect sizes should be heterogeneous across studies in the vast majority of cases (see, for example, National Research Council, 1992), and significance tests of this variability have low power. Therefore, variability statistics should be reported, regardless of whether moderator variables have been measured, because they tell us something important about the distribution of effect sizes in the meta-analysis, but not as a justification for choosing a particular method.
(3) *Moderator variables*. If there is variability in effect sizes, and in most cases there is (Field, 2005b), this variability can be explored in terms of moderator variables (Field, 2003b; Overton, 1998). For example, we might find that compared to a waiting-list control, CBT including group therapy produces a larger effect size for improvement in adolescents with eating disorders than CBT without a group component.

This article is intended as an extended tutorial in which we overview the key stages necessary when conducting a meta-analysis. The article describes how to do meta-analysis in a step-by-step way using some examples from the psychological literature. In doing so, we look not only at the theory of meta-analysis, but also at how to use computer programs to conduct one: we focus on IBM SPSS, because many psychologists use it, and R (because it is free and does things that SPSS cannot). We have broken the process of meta-analysis into six steps: (1) do a literature search; (2) decide on some inclusion criteria and apply them; (3) calculate effect sizes for each study to be included; (4) do the basic meta-analysis; (5) consider doing some more advanced analysis such as publication bias analysis and exploring moderator variables; and (6) write up the results.

The example data sets

In this tutorial, we use two real data sets from the psychological literature. Cartwright-Hatton *et al.* (2004) conducted a systematic review of the efficacy of CBT for childhood and adolescent anxiety. This study is representative of clinical research in that relatively few studies had addressed this question and sample sizes within each study were relatively small. These data are used as our main example and the most benefit can be gained from reading their paper in conjunction with this one. When discussing

moderator analysis, we use a larger data set from Tenenbaum and Leaper (2002), who conducted a meta-analysis on whether parents' gender schemas related to their children's gender-related cognitions. These data files are available on the website that accompanies this article (Field & Gillett, 2009).

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, using electronic databases such as the ISI Web of Knowledge, PubMed and PsycINFO. This can be done to find articles, but also to identify authors in the field (who might have unpublished data – see below); in the later case, it can be helpful not only to backward-search for articles but also to forward-search by finding authors who cite papers in the field. It is often useful to hand-search relevant journals that are not part of these electronic databases and to use the reference sections of the articles that you have found to check for articles that you have missed. One potential bias in a meta-analysis arises from the fact that significant findings are more likely to be published than non-significant findings because researchers do not submit them (Dickersin, Min, & Meinert, 1992) and reviewers tend to reject manuscripts containing them (Hedges, 1984). This is known as publication bias or the 'file-drawer' problem (Rosenthal, 1979). This bias is not trivial: significant findings are estimated to be eight times more likely to be submitted than non-significant ones (Greenwald, 1975), studies with positive findings are around seven times more likely to be published than studies with results supporting the null hypothesis (Coursol & Wagner, 1986), and 97% of articles in psychology journals report significant results (Sterling, 1959). The effect of this bias is that meta-analytic reviews will overestimate population effects if they have not included unpublished studies, because effect sizes in unpublished studies of comparable methodological quality will be smaller (McLeod & Weisz, 2004) and can be half the size of comparable published research (Shadish, 1992). To minimize the bias of the file-drawer problem, the search can be extended from papers to relevant conference proceedings, and by contacting people whom you consider to be experts in the field to see if they have any unpublished data or know of any data relevant to your research question that is not in the public domain. This can be done by direct e-mail to authors in the field, but also by posting a message to a topic specific newsgroup or using LISTSERV.

Turning to our example, Cartwright-Hatton *et al.* (2004) gathered articles by searching eight databases: Cochrane Controlled Trials register, Current Controlled Trials, Medline, Embase/PsycINFO, Cinahl, NHS Economic Evaluation Database, National Technical Information Service, and ISI Web of Science. They also searched the reference lists of these articles, and hand-searched 13 journals known to publish clinical trials on anxiety or anxiety research generally. Finally, the authors contacted people in the field and requested information about any other trials not unearthed by their search. This search strategy highlights the use of varied resources to ensure all potentially relevant studies are included and to reduce bias due to the file-drawer problem.

Step 2: Decide on inclusion criteria

The inclusion of badly conducted research can also bias a meta-analysis. Although meta-analysis might seem to solve the problem of variance in study quality because these differences will 'come out in the wash', even one red sock (bad study) amongst the

white clothes (good studies) can ruin the laundry. Meta-analysis can end up being an exercise in adding apples to oranges unless inclusion criteria are applied to ensure the quality and similarity of the included studies.

Inclusion criteria depend on the research question being addressed and any specific methodological issues in the field; for example, in a meta-analysis of a therapeutic intervention such as CBT, you might decide on a working definition of what constitutes CBT, and perhaps exclude studies that do not have proper control groups and so on. You should not exclude studies because of some idiosyncratic whim: it is important that you formulate a precise set of criteria that is applied throughout; otherwise you will introduce subjective bias into the analysis. It is also vital to be transparent about the criteria in your write-up, and even consider reporting the number of studies that were included/excluded at each hurdle in the process.

It is also possible to classify studies into groups, for example methodologically strong or weak, or the use of waiting-list controls or other intervention controls, and then see if this variable moderates the effect size; by doing so you can answer questions such as: do methodologically strong studies (by your criteria) differ in effect size from the weaker studies? Or, does the type of control group affect the strength of the effect of CBT?

The Cartwright-Hatton *et al.* (2004) review lists a variety of inclusion criteria that will not be repeated here; reading their paper though will highlight the central point that they devised criteria sensible to their research question: they were interested in child anxiety, so variables such as age of patients (were they children?), diagnostic status (were they anxious?), and outcome measures (did they meet the required standard?) were used as inclusion criteria.

Step 3: Calculate the effect sizes

What are effect sizes and how do I calculate them?

Once you have collected your articles, you need to find the effect sizes within them, or calculate them for yourself. An effect size is usually a standardized measure of the magnitude of observed effect (see, for example, Clark-Carter, 2003; Field, 2005c). As such, effect sizes across different studies that have measured different variables, or have used different scales of measurement, can be directly compared: an effect size based on the Beck Anxiety Inventory could be compared to an effect size based on heart rate. Many measures of effect size have been proposed (see Rosenthal, 1991, for a good overview) and the most common are Pearson's correlation coefficient, r, Cohen's d, and the odds ratio (OR). However, there may be reasons to prefer unstandardized effect size measures (Baguley, 2009), and meta-analytic methods exist for analysing these that will not be discussed in this paper (but see Bond, Wiitala, & Richard, 2003).

Pearson's correlation coefficient, r, is a standardized form of the covariance between two variables and is well known and understood by most psychologists as a measure of the strength of relationship between two continuous variables; however, it is also a very versatile measure of the strength of an experimental effect. If you had a sports-related concussion group (coded numerically as 1) and a non-concussed control (coded numerically as 0), and you conducted a Pearson correlation between this variable and their performance on some cognitive task, the resulting correlation will have the same p value as a t test on the same data. In fact, there are direct relationships between r and statistics that quantify group differences (e.g., t and F), associations between categorical

variables (χ^2), and the *p* value of any test statistic. The conversions between *r* and these various measures are discussed in many sources (e.g., Field, 2005a, 2005c; Rosenthal, 1991) and will not be repeated here.

Cohen (1988, 1992) made some widely adopted suggestions about what constitutes a large or small effect: $r = .10$ (small effect, explaining 1% of the total variance); $r = .30$ (medium effect, accounting for 9% of the total variance); $r = .50$ (large effect, accounting for 25% of the variance). Although these guidelines can be a useful rule of thumb to assess the importance of an effect (regardless of the significance of the test statistic), it is worth remembering that these 'canned' effect sizes are not always comparable when converted to different metrics, and that there is no substitute for evaluating an effect size within the context of the research domain in which it is being used (Baguley, 2009; Lenth, 2001).

Cohen's *d* is based on the standardized difference between two means. You subtract the mean of one group from the mean of the other and then standardize this by dividing by σ, which is the sum of squared errors (i.e., take the difference between each score and the mean, square it, and then add all of these squared values up) divided by the total number of scores:

$$d = \frac{M_1 - M_2}{\sigma}.$$

σ either can be based on a single group (usually the control group) or can be a pooled estimate based on both groups by using the sample size, *n*, and variances, *s*, from each:

$$\sqrt{\frac{(n_1 - 1)s_1 + (n_2 - 1)s_2}{n_1 + n_2 - 2}}.$$

Whether you standardize using one group or both depends on what you are trying to quantify. For example, in a clinical drug trial, the drug dosage will affect not just the mean of any outcome variables, but also the variance; therefore, you would not want to use this inflated variance when computing *d* and would instead use the control group only (so that *d* reflects the mean change relative to the control group).

If some of the primary studies have employed factorial designs, it is possible to obtain estimators of effect size for these designs that are metrically comparable with the *d* estimator for the two-group design (Gillett, 2003). As with *r*, Cohen (1988, 1992) has suggested benchmarks of $d = .30$, .50 and .80 as representing small, medium and large effects, respectively.

The OR is the ratio of the odds (the probability of the event occurring divided by the probability of the event not occurring) of an event occurring in one group compared to another (see Fleiss, 1973). For example, if the odds of being symptom-free after treatment are 10, and the odds of being symptom-free after being on the waiting list are 2 then the OR is $10/2 = 5$. This means that the odds of being symptom-free are five times greater after treatment, compared to being on the waiting list. The OR can vary from 0 to infinity, and a value of 1 indicates that the odds of a particular outcome are equal in both groups. If dichotomized data (i.e., a 2×2 contingency table) need to be incorporated into an analysis based mainly on *d* or *r*, then a *d*-based measure called d_{Cox} exists (see Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003, for a review).

There is much to recommend r as an effect size measure (e.g., Rosenthal & DiMatteo, 2001). It is certainly convenient because it is well understood by most psychologists, and unlike d and the OR it is constrained to lie between 0 (no effect) and ± 1 (a perfect effect). It does not matter what effect you are looking for, what variables have been measured, or how those variables have been measured: a correlation coefficient of 0 means there is no effect, and a value of ± 1 means that there is a perfect association. (Note that because r is not measured on a linear scale, an effect such as $r = .4$ is not twice as big as one with $r = .2$.) However, there are situations in which d may be favoured; for example, when group sizes are very discrepant (McGrath & Meyer, 2006) r might be quite biased because, unlike d, it does not account for these 'base rate' differences in group n. In such circumstances, if r is used it should be adjusted to the same underlying base rate, which could be the base rate suggested in the literature, the average base rate across studies in the meta-analysis, or a 50/50 base rate (which maximizes the correlation).

Whichever effect size metric you chose to use, your next step will be to go through the articles that you have chosen to include and calculate effect sizes using your chosen metric for comparable effects within each study. If you were using r, this would mean obtaining a value for r for each effect that you wanted to compare for every paper you want to include in the meta-analysis. A given paper may contain several rs depending on the sorts of questions you are trying to address with your meta-analysis. For example, cognitive impairment in post-traumatic stress disorder could be measured in a variety of ways in individual studies and so a meta-analysis might use several effect sizes from the same study (Brewin, Kleiner, Vasterling, & Field, 2007). Solutions include calculating the average effect size across all measures of the same outcome within a study (Rosenthal, 1991), comparing the meta-analytic results when allowing multiple effect sizes from different measures of the same outcome within a study, or computing an average effect size so that every study contributes only one effect to the analysis (as in Brewin *et al.*, 2007).

Articles might not report effect sizes, or might report them in different metrics. If no effect sizes are reported then you can often use the reported data to calculate one. For most effect size measures, you could do this using test statistics (as mentioned above, r can be obtained from t, z, χ^2, and F), or probability values for effects (by converting first to z). If you use d as your effect size then you can use means and standard deviations reported in the paper. Finally, if you are calculating ORs then frequency data from the paper could be used. Sometimes papers do not include sufficient data to calculate an effect size, in which case contact the authors for the raw data, or relevant statistics from which an effect size can be computed. (Such attempts are often unsuccessful and we urge authors to be sympathetic to e-mails from meta-analysts trying to find effect sizes.) If a paper reports an effect size in a different metric than the one that you have chosen to use then you can usually convert from one metric to another to at least get an approximate effect size.[1] A full description of the various conversions is beyond the scope of this article, but many of the relevant equations can be found in Rosenthal (1991). There are also many Excel spreadsheets

[1] These conversions are often approximate and can have statistical implications. However, r can be converted to d approximately using the equation: $r = d/\sqrt{d^2 + 1/pq}$, in which p is the proportion of participants in the first group and q is the proportion of participants in the second group. To convert the opposite way (again this conversion is approximate), use $d = r/\sqrt{pq(1 - r^2)}$. Similarly, r can be obtained from an OR: $r = (\mathrm{OR}^{0.5} - 1)/(\mathrm{OR}^{0.5} + 1)$.

available on-line that compute effect sizes and convert between them; some examples are DeCoster (1998) and Wilson (2004).

Calculating effect sizes for Cartwright-Hatton et al. *(2004)*

When reporting a meta-analysis it is a good idea to tabulate the effect sizes with other helpful information (such as the sample size on which the effect size is based, *N*) and also to present a stem-and-leaf plot of the effect sizes. For the Cartwright-Hatton *et al.* data, we used *r* as the effect size measure but we will highlight differences for situations in which *d* is used when we talk about the meta-analysis itself. Table 1 shows a stem-and-leaf plot of the resulting effect sizes and this should be included in the write-up. This stem-and-leaf plot tells us the exact effect sizes to two decimal places, with the stem reflecting the first decimal place and the leaf showing the second;

Table 1. Stem-and-leaf plot of all effect sizes (*r*s)

Stem	Leaf
.0	
.1	8
.2	
.3	
.4	
.5	0, 5, 8
.6	0, 2, 5
.7	1, 2
.8	5
.9	

for example, we know the smallest effect size was $r = .18$, the largest was $r = .85$, and there were effect sizes of .71 and .72. Table 2 shows the studies included in Cartwright-Hatton *et al.* (2004), with their corresponding effect sizes (expressed as *r*) and the sample sizes on which these *r*s are based.

Table 2. Calculating the Hunter–Schmidt estimate

Study	*N*	*r*	$N \times r$
Barrett (1998)	50	.55	27.30
Barrett *et al.* (1996)	76	.50	38.12
Dadds *et al.* (1997)	93	.18	16.49
Flannery-Schroeder and Kendall (2000)	43	.85	36.45
Hayward *et al.* (2000)	33	.62	20.43
Kendall (1994)	45	.71	31.85
Kendall *et al.* (1997)	70	.58	40.90
Shortt *et al.* (2001)	65	.65	42.03
Silverman *et al.* (1999)	41	.60	24.69
Spence *et al.* (2000)	47	.72	33.80
Total	563		312.06

Step 4: Do the basic meta-analysis

Having collected the relevant studies and calculated effect sizes from each study, you must do the meta-analysis. This section looks first at some important conceptual issues before exploring how to actually do the meta-analysis.

Initial considerations

The main function of meta-analysis is to estimate effects in the population 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, which is typically a function of sample size.[2] This makes statistical sense because if an effect size has good sampling accuracy (i.e., it is likely to be an accurate reflection of reality) then it is weighted highly, whereas effect sizes that are imprecise are given less weight in the calculations. It is usually helpful to also construct a confidence interval around the estimate of the population effect. Data analysis is rarely straightforward, and meta-analysis is no exception because there are different methods for estimating the population effects and these methods have their own pros and cons. There are lots of issues to bear in mind and many authors have written extensively about them (Field, 2001, 2003a, 2003b, 2005b, 2005c; Hall & Brannick, 2002; Hunter & Schmidt, 2004; Rosenthal & DiMatteo, 2001; Schulze, 2004). In terms of doing a meta-analysis, the main issues (as we see them) are: (1) which method to use and (2) how to conceptualize your data. Actually, these two issues are linked.

Which method should I chose?

In essence, there are two ways to conceptualize meta-analysis: fixed- and random-effects models (Hedges, 1992; Hedges & Vevea, 1998; Hunter & Schmidt, 2000).[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 or can be predicted from a few predictors (Hunter & Schmidt, 2000). Consequently, sample effect sizes should be homogeneous 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 'superpopulation' (Hedges, 1992). In this case, the effect sizes should be heterogeneous because they come from populations with varying average effect sizes.

The above distinction is tied up with the method of meta-analysis that you chose because statistically speaking the main difference between fixed- and random-effects models is in the sources of error. In fixed-effects models, there is error 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. As such, calculating the error of the mean effect size in random-effects

[2] If there is a study with a hugely discrepant sample size (i.e., very large compared to other studies) you should consider conducting the analysis with and without this study to assess the extent to which this study will bias the results.

[3] A mixed-effects model exists too in which population effect sizes differ but their variability is explained by a moderator variable that is treated as 'fixed' (see Overton, 1998) and also includes additional random heterogeneity.

models involves estimating two error terms, whereas in fixed-effects models there is only one. This, as we will see, has implications for computing the mean effect size. The two most widely used methods of meta-analysis are those by Hunter and Schmidt (2004) which is a random-effects method, and by Hedges and colleagues (e.g., Hedges, 1992; Hedges & Olkin, 1985; Hedges & Vevea, 1998) who provide both fixed- and random-effects methods. However, multi-level models can also be used in the context of meta-analysis (see Hox, 2002, chap. 8).

Before doing the actual meta-analysis, you need to decide whether to conceptualize your model as fixed or random effects. This decision depends both on the assumptions that can realistically be made about the populations from which your studies are sampled, and the types of inferences that you wish to make from the meta-analysis. On the former point, many writers have argued that real-world data in the social sciences are likely to have variable population parameters (Field, 2003b; Hunter & Schmidt, 2000, 2004; National Research Council, 1992; Osburn & Callender, 1992). There are data to support these claims: Field (2005b) calculated the standard deviations of effect sizes for all meta-analytic studies (using *r*) published in *Psychological Bulletin* in 1997–2002 and found that they ranged from 0 to 0.3, and were most frequently in the region of 0.10–0.16; Barrick and Mount (1991) similarly found that the standard deviation of effect sizes (*r*s) in published data sets was around 0.16. These data suggest that a random-effects approach should be the norm in social science data.

The decision to use fixed- or random-effects models also depends upon the type of inferences that you wish to make (Hedges & Vevea, 1998): fixed-effect models are appropriate for inferences that extend only to the studies included in the meta-analysis (conditional inferences), whereas random-effects models allow inferences that generalize beyond the studies included in the meta-analysis (unconditional inferences). Psychologists will typically wish to generalize their findings beyond the studies included in the meta-analysis and so a random-effects model is appropriate.

The decision about whether to apply fixed- or random-effects methods is not trivial. Despite considerable evidence that variable effect sizes are the norm in psychological data, fixed-effects methods are routinely used: a review of meta-analytic studies in *Psychological Bulletin* found 21 studies using fixed-effects methods (in 17 of these studies there was significant variability in sample effect sizes) and none using random-effects methods (Hunter & Schmidt, 2000). The consequences of applying fixed-effects methods to random-effects data can be quite dramatic: significance tests of the estimate of the population effect have Type I error rates inflated from the normal 5% to 11–28% (Hunter & Schmidt, 2000) or 43–80% (Field, 2003b), depending on the variability of effect sizes. In addition, when applying two random-effects methods to 68 meta-analyses from five large meta-analytic studies published in *Psychological Bulletin*, Schmidt, Oh, and Hayes (2009) found that the published fixed-effects confidence intervals around mean effect sizes were on average 52% narrower than their actual width: these nominal 95% fixed-effects confidence intervals were on average 56% confidence intervals. The consequences of applying random-effects methods to fixed-effects data are considerably less dramatic: in Hedges' method, for example, the additional between-study effect size variance used in the random-effects method becomes zero when sample effect sizes are homogeneous, yielding the same result as the fixed-effects method.

We mentioned earlier that part of conducting a meta-analysis is to compute statistics that quantify heterogeneity. These tests can be used to ascertain whether population effect sizes are likely to be fixed or variable (Hedges & Olkin, 1985). If these homogeneity tests yield non-significant results then sample effect sizes are usually

regarded as roughly equivalent and so population effect sizes are likely to be homogeneous (and hence the assumption that they are fixed is reasonable). However, these tests should be used cautiously as a means to decide on how to conceptualize that data because they typically have low power to detect genuine variation in population effect sizes (Hedges & Pigott, 2001). In general, we favour the view that the choice of model should be determined *a priori* by the goal of the analysis rather than being a *post hoc* decision based on the data collected.

To sum up, we believe that in most cases a random-effects model should be assumed (and the consequences of applying random-effects models to fixed-effects data are much less severe than the other way around). However, fixed-effects analysis may be appropriate when you do not wish to generalize beyond the effect sizes in your analysis (Oswald & McCloy, 2003); for example, a researcher who has conducted several similar studies some of which were more successful than others might reasonably estimate the population effect of her research by using a fixed-effects analysis. For one thing, it would be reasonable for her to assume that her studies are tapping the same population, and also, she would not necessarily be trying to generalize beyond her own studies.

Which method is best?

The next decision is whether to use Hunter and Schmidt (2004) and Hedges and colleagues' method.[4] These methods will be described in due course, and the technical differences between them have been summarized by Field (2005b) and will not be repeated here. Field (2001; but see Hafdahl & Williams, 2009) conducted a series of Monte Carlo simulations comparing the performance of the Hunter and Schmidt and Hedges and Olkin (fixed and random effects) methods and found that when comparing random-effects methods the Hunter–Schmidt method yielded the most accurate estimates of population correlation across a variety of situations – a view echoed by Hall and Brannick (2002) in a similar study. However, neither the Hunter–Schmidt nor Hedges and colleagues' method controlled the Type I error rate when 15 or fewer studies were included in the meta-analysis, and the method described by Hedges and Vevea (1998) controlled the Type I error rate better than the Hunter–Schmidt method when 20 or more studies were included. Schulze (2004) has also done extensive simulation studies and, based on these findings, recommends against using Fisher's z transform and suggests that the 'optimal' study weights used in the Hedges–Vevea method can, at times, be suboptimal in practice. However, Schulze based these conclusions on using only the fixed-effects version of Hedges' method. Field (2005b) looked at Hedges and colleagues' random-effects method and again compared it to Hunter and Schmidt's bare-bones method using a Monte Carlo simulation. He concluded that in general both random-effects methods produce accurate estimates of the population effect size. Hedges' method showed small (less than .052 above the population correlation) overestimations of the population correlation in extreme situations (i.e., when the population correlation was large, $\bar{\rho} \geq .3$, and the standard deviation of correlations was also large, $\sigma_\rho \geq 0.16$; also when the population correlation was small, $\bar{\rho} \geq .1$ and the standard deviation of correlations was at its maximum value,

[4] There are other methods. For example, Rosenthal and Rubin's (1978) method is a fixed-effect method and differs from Hedges' method only in how the significance of the mean-weighted effect size is calculated (see Field, 2001). Given that significance testing of the population effect estimate should not be the key concern in meta-analysis, we have omitted this method.

$\sigma_\rho = 0.32$). The Hunter–Schmidt estimates were generally less biased than estimates from Hedges' random-effects method (less than .011 below the population value), but in practical terms the bias in both methods was negligible. In terms of 95% confidence intervals around the population estimate, Hedges' method was in general better at achieving these intervals (the intervals for Hunter and Schmidt's method tended to be too narrow, probably because they recommend using credibility intervals and not confidence intervals – see below). However, the relative merits of the methods depended on the parameters of the simulation and in practice the researcher should consult the various tables in Field (2005b) to assess which method might be most accurate for the given parameters of the meta-analysis that they are about to conduct. Also, Hunter and Schmidt's method involves psychometric corrections for the attenuation of observed effect sizes that can be caused by measurement error (Hunter, Schmidt, & Le, 2006). Not all studies will report reliability coefficients, so their methods use the average reliability across studies to correct effect sizes. These psychometric corrections can be incorporated into any procedure, including that of Hedges and colleagues, but these conditions are not explored in the comparison studies mentioned above.

Methods of meta-analysis

Hunter–Schmidt method

As already mentioned, this method emphasizes isolating and correcting for sources of error such as sampling error and reliability of measurement variables. However, Hunter and Schmidt (2004) spend an entire book explaining these corrections, and so for this primer we will conduct the analysis in its simplest form. The population effect is estimated using a simple mean in which each effect size estimate, r, is weighted by the sample size on which it is based, n:

$$\bar{r} = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i}. \tag{1}$$

Table 2 shows the effect sizes and their sample sizes, and in the final column we have multiplied each effect size by the sample size on which it is based. The sum of this final column is the numerator of equation (1), whereas the sum of sample sizes (column 2 in Table 2) is the denominator of this equation. Therefore, the population effect can be estimated as

$$\bar{r} = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i} = \frac{312.06}{563.00} = .554.$$

By Cohen's (1988, 1992) criteria, this means that CBT for childhood and adolescent anxiety had a large effect compared to waiting-list controls.

The next step is to estimate the generalizability of this value using a credibility interval.[5] Hunter and Schmidt (2004) recommend correcting the population effect for

[5] Credibility intervals differ from confidence intervals. In essence, confidence intervals measure the precision of an estimate, whereas credibility intervals reflect whether validity can be generalized. For example, CBT may be an effective therapy for children with panic disorder, but not for children with social anxiety. As such, credibility intervals address whether other variables moderate the population effect, or whether the population of effect sizes should be broken down into subpopulations (Whitener, 1990). In contrast, confidence intervals indicate the effect that sampling error has had on the estimate of the population effect.

artefacts before constructing these credibility intervals. If we ignore artefact correction, the credibility intervals are based on the variance of effect sizes in the population. 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:

$$\hat{\sigma}_r^2 = \frac{\sum_{i=1}^{k} n_i(r_i - \bar{r})^2}{\sum_{i=1}^{k} n_i}. \tag{2}$$

It is also necessary to estimate the sampling error variance using the population correlation estimate, $\bar{r}$, and the average sample size, $\bar{N}$ (see Hunter & Schmidt, 2004, p. 88):

$$\hat{\sigma}_e^2 = \frac{(1 - \bar{r}^2)^2}{\bar{N} - 1}. \tag{3}$$

To estimate the variance in population correlations, we subtract the sampling error variance from the variance in sample correlations (see Hunter & Schmidt, 2004, p. 88):

$$\hat{\sigma}_\rho^2 = \hat{\sigma}_r^2 - \hat{\sigma}_e^2. \tag{4}$$

The credibility intervals are based on taking the population effect estimate (equation (1)) and adding to or subtracting from it the square root of the estimated population variance in equation (4) multiplied by $z_{\alpha/2}$, in which α is the desired probability (e.g., for a 95% interval, $z_{\alpha/2} = 1.96$):

$$\begin{aligned} 95\%\,\text{credibility interval}_{\text{upper}} &= \bar{r} + 1.96\sqrt{\hat{\sigma}_\rho^2}, \\ 95\%\,\text{credibility interval}_{\text{lower}} &= \bar{r} - 1.96\sqrt{\hat{\sigma}_\rho^2}. \end{aligned} \tag{5}$$

A chi-square statistic is used to measure homogeneity of effect sizes. This statistic is based on the sum of squared errors of the mean effect size, and it is calculated from the sample size on which the correlation is based (n), the squared errors between each effect size and the mean, and the variance:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - 1)(r_i - \bar{r})^2}{(1 - \bar{r}^2)^2}. \tag{6}$$

Hedges and colleagues' method

In this method (Hedges & Olkin, 1985; Hedges & Vevea, 1998), if r is being used, effect sizes are first converted into a standard normal metric, using Fisher's (1921) r-to-z transformation, before calculating a weighted average of these transformed scores (in which r_i is the effect size from study i). Fisher's transformation is given by

$$z_{r_i} = \frac{1}{2}\log_e\left(\frac{1 + r_i}{1 - r_i}\right), \tag{7}$$

and the reverse transformation by

$$r_i = \frac{e^{2z_{r_i}} - 1}{e^{2z_{r_i}} + 1} \tag{8}$$

To remove the slight positive bias found from Fisher-transformed *r*s, the effect sizes can be transformed with $r - [(r(1 - r^2))/2(n - 3)]$ before the Fisher transformation in equation (7) is applied (see Overton, 1998). This is done in the SPSS syntax files that we have produced to accompany this paper. Note also that less biased *r*-to-*z* transformations have been developed that may explain some of the differences between the two methods of meta-analysis discussed in this paper (Hafdahl, 2009, 2010).

In the fixed-effects 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.

$$\bar{z}_r = \frac{\sum_{i=1}^{k} w_i z_{r_i}}{\sum_{i=1}^{k} w_i}, \tag{9}$$

in which k is the number of studies in the meta-analysis. When r is the effect size measure, the weight (w_i) is the sample size, n_i, less 3 ($w_i = n_i - 3$), but when d is the effect size measure this weight is $w_i = 4N_i(1 + d_i^2)/8$. The resulting weighted average is in the z-metric and should be converted back to r using equation (8).

We use this average, and the weight for each study, to calculate the homogeneity of effect sizes. The resulting statistic Q has a chi-square distribution with $k - 1$ degrees of freedom:

$$Q = \sum_{i=1}^{k} w_i \left(z_{r_i} - \bar{z}_r\right)^2. \tag{10}$$

If you wanted to apply a fixed-effects model you could stop here. However, as we have suggested, there is usually good reason to assume that a random-effects model is most appropriate. To calculate the random-effects average effect size, the weights use a variance component that incorporates both between- and within-study variance. The between-study variance is denoted by τ^2 and is added to the within-study variance to create new weights:

$$w_i^* = \left(\frac{1}{w_i} + \hat{\tau}^2\right)^{-1}. \tag{11}$$

The value of w_i depends upon whether r or d has been used (see above): when r has been used, $w_i = n_i - 3$. The random-effects weighted average in the z metric uses the same equation as the fixed-effects model, except that the weights have changed to incorporate between-study variance:

$$\bar{z}_r^* = \frac{\sum_{i=1}^{k} w_i^* z_{r_i}}{\sum_{i=1}^{k} w_i^*}. \tag{12}$$

The between-studies variance can be estimated in several ways (see Friedman, 1937; Hedges & Vevea, 1998; Overton, 1998; Takkouche, Cadarso-Suárez, & Spiegelman, 1999).

Hedges and Vevea (1998, equation (10)) use an equation based on Q (the weighted sum of squared errors in equation (10)), k, and a constant, c,

$$\hat{\tau}^2 = \frac{Q - (k - 1)}{c}, \tag{13}$$

where c is defined as

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} (w_i)^2}{\sum_{i=1}^{k} w_i}. \tag{14}$$

If the estimate of between-studies variance, $\hat{\tau}^2$, yields a negative value then it is set to zero (because the variance between studies cannot be negative). The estimate $\hat{\tau}^2$ is substituted in equation (11) to calculate the weight for a particular study, and this in turn is used in equation (12) to calculate the average correlation. This average correlation is then converted back to the r metric using equation (8) before being reported.

The final step is to estimate the precision of this population effect estimate using confidence intervals. The confidence interval for a mean value is calculated using the standard error of that mean. Therefore, to calculate the confidence interval for the population effect estimate, we need to know the standard error of the mean effect size. This is the square root of the reciprocal of the sum of the random-effects weights (see Hedges & Vevea, 1998, p. 493):

$$SE(\bar{z}_r^*) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}}. \tag{15}$$

The confidence interval around the population effect estimate is calculated in the usual way by multiplying the standard error by the two-tailed critical value of the normal distribution (which is 1.96 for a 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:

$$\begin{aligned} 95\%\,\mathrm{CI}_{\mathrm{upper}} &= \bar{z}_r^* + 1.96SE(\bar{Z}_r^*), \\ 95\%\,\mathrm{CI}_{\mathrm{lower}} &= \bar{z}_r^* - 1.96SE(\bar{Z}_r^*). \end{aligned} \tag{16}$$

These values are again transformed back to the r metric using equation (8) before being reported.

Doing meta-analysis on a computer

In reality, you will not do the meta-analysis by hand (although we believe that there is no harm in understanding what is going on behind the scenes). There are some stand-alone packages for conducting meta-analyses such as Comprehensive Meta-Analysis, which implements many different meta-analysis methods, converts effect sizes, and creates plots of study effects. Hunter and Schmidt (2004) provide specialist custom-written software for implementing their full method on the CD-ROM of their book. There is also a program called Mix (Bax, Yu, Ikeda, Tsuruta, & Moons, 2006), and the Cochrane

Collaboration (2008) provides software called Review Manager for conducting meta-analysis. Both of these packages have excellent graphical facilities.

For those who want to conduct meta-analysis without the expense of buying specialist software, meta-analysis can also be done using R (R Development Core Team, 2008), a freely available package for conducting a staggering array of statistical procedures. R is based on the S language and so has much in common with the commercially available package S-PLUS. Scripts for running a variety of meta-analysis procedures on *d* are available in the 'meta' package that can be installed into R (Schwarzer, 2005). Likewise, publication bias analysis can be run in R. The implementation of some of these programs will be described in due course. In addition, Excel users can use a plug-in called MetaEasy (Kontopantelis & Reeves, 2009).

SPSS does not, at present, offer built-in tools for doing meta-analysis, but the methods described in this paper can be conducted using custom-written syntax. To accompany this article, we have produced syntax files to conduct many of the meta-analytic techniques discussed in this paper (although the Hunter–Schmidt version is in only its simplest form). Other SPSS syntax files for *r* and also *d* are also available from Lavesque (2001) and Wilson (2004). All of the data and syntax files accompanying this paper can be downloaded from our webpage (Field & Gillett, 2009). The SPSS syntax files are:

(1) *Basic meta-analysis*. The files Meta_Basic_r.sps, Meta_Basic_d.sps, and Meta_Basic_D_h.sps can be used to perform a basic meta-analysis on effect sizes expressed as *r*, *d*, and the difference between proportions (*D* or *h*), respectively, in SPSS. The output provides an estimate of the average effect size of all studies, or any subset of studies, a test of homogeneity of effect size that contributes to the assessment of the goodness of fit of the statistical model, elementary indicators of, and tests for, the presence of publication bias, and parameters for both fixed- and random-effects models.

(2) *Moderator variable analysis*. The files Meta_Mod_r.sps, Meta_Mod_d.sps, and Meta_Mod_D_h.sps can be used for analysing the influence of moderator variables on effect sizes expressed as *r*, *d*, and the difference between proportions (*D* or *h*), respectively, in SPSS. Each of these files is run using a shorter syntax file to launch these files (i.e., Meta_Mod_r.sps is launched by using the syntax file Launch_Meta_Mod_r.sps). The programs use weighted multiple regression to provide an evaluation of the impact of continuous moderator variables on effect sizes, an evaluation of the impact of categorical moderator variables on effect sizes, tests of homogeneity of effect sizes that contribute to the assessment of the goodness of fit of a statistical model incorporating a given set of moderator variables, and estimates for both fixed- and random-effects models.

(3) *Publication bias analysis*. The files Pub_Bias_r.R, Pub_Bias_d.R, and Pub_Bias_D_h.R can be used to produce funnel plots and a more sophisticated publication bias analysis on effect sizes expressed as *r*, *d*, and the difference between proportions (*D* or *h*), respectively, using the software R. Each file computes an ordinary unadjusted estimate and four adjusted estimates of effect size that indicate the potential impact of severe and moderate one- and two-tailed bias, for both fixed- and random-effects models (see below). Note that these files include Vevea and Woods's (2005) scripts for R.

In this section, we will do the basic analysis that we described in the previous section using the effect sizes from Cartwright-Hatton *et al.* (2004) expressed as *r*. Before we begin, you need to create a folder in the 'Documents' folder on your hard drive called 'Meta-Analysis' (for the first author, the complete file path would, therefore, be 'C:\Users\Dr. Andy Field\Documents\Meta-Analysis'.[6] This folder is needed for some of our files to work.

In the SPSS data editor, create two new variables, the first for the effect sizes, *r*, and the second for the total sample size on which each effect size is based, *n*; it is also good practice to create a variable in which you identify the study from which each effect size came. You can download this data file (Cartwright-Hatton_et_al_2004.sav) from the accompanying website. Once the data are entered, simply open the syntax file and in the syntax window click on the 'Run' menu and then select 'All'. The resulting output is in Figure 1. Note that the analysis calculated both fixed- and random-effects statistics for both methods. This is for convenience but, given that we have made an *a priori* decision about which method to use, and whether to apply a fixed- and random-effects analysis, we would interpret only the corresponding part of the output. In this case, we opted for a random-effects analysis. This output is fairly self-explanatory; for example, we can see that, for Hedges and Vevea's method, the Q statistic (equation (10) above) is highly significant, $\chi^2(9) = 41.27$, $p < .001$. Likewise, the population effect size once returned to the *r* metric and its 95% confidence interval are: .61 (95% CI [.48, .72]). We can also see that this population effect size is significant, $z = 7.57$, $p < .001$.

At the bottom of the output are the corresponding statistics from the Hunter–Schmidt method including the population estimate, .55, the sample correlation variance from equation (2), .036, the sampling error variance from equation (3), .009, the variance in population correlations from equation (4), .027, the upper and lower bounds of the credibility interval from equation (5), .87 and .23, and the chi-square test of homogeneity from equation (6) and its associated significance, $\chi^2(9) = 41.72$, $p < .001$. The output also contains important information to be used to estimate the effects of publication bias, but we will come back to this issue in due course.

Based on both homogeneity tests, we could say that there was considerable variation in effect sizes overall. Also, based on the estimate of population effect size and its confidence interval, we could conclude that there was a strong effect of CBT for childhood and adolescent anxiety disorders compared to waiting-list controls. To get some ideas about how to write up a meta-analysis like this, see Brewin *et al.* (2007).

Step 5: Do some more advanced analysis

Moderator analysis

Theory behind moderator analysis

The model for moderator effects is a mixed model (which we mentioned earlier): it assumes a general linear model in which each *z*-transformed effect size can be

[6] Our files are written for Windows (Vista onwards). However if you use an earlier version simply edit the line 'cd "%HOMEDRIVE%%HOMEPATH%\Documents\Meta-Analysis"' to include 'My Documents' (i.e., 'cd "%HOMEDRIVE%%HOMEPATH%\My Documents\Meta-Analysis"'.) Mac users should replace this line of code with 'cd "Documents\Meta-Analysis"'.

```
Run MATRIX procedure:

**********   META-ANALYSIS OF CORRELATION COEFFICIENTS:  r   **********

NUMBER OF STUDIES
         k
        10

**********   FIXED-EFFECTS MODEL   **********

MEAN EFFECT SIZE, LOWER & UPPER 95% CONFIDENCE BOUNDS, AND Z-TEST
    Mean r   Lower r   Upper r          z          p          k
      .580       .520       .633    15.278       .000    10.000

HOMOGENEITY TEST:  Q STATISTIC  (Goodness of Fit)
      Chi2         df          p
    41.268     9.000       .000

**********   HEDGES-VEVEA RANDOM-EFFECTS MODEL   **********

MEAN EFFECT SIZE, LOWER & UPPER 95% CONFIDENCE BOUNDS, AND Z-TEST
    Mean r   Lower r   Upper r          z          p          k
      .612       .484       .715     7.575       .000    10.000

Estimated Variance in Population (Fisher-Transformed) Correlations
       Tau
     .0681

HOMOGENEITY TEST:  Q STATISTIC  (Goodness of Fit)
      Chi2         df          p
     7.891     9.000       .545

**********   HUNTER-SCHMIDT RANDOM-EFFECTS MODEL   **********

MEAN EFFECT SIZE, LOWER & UPPER 95% CREDIBILITY BOUNDS, AND CHI-SQUARE TEST
    Mean r   Lower r   Upper r       Chi2          p         df
      .551       .229       .873    40.897       .000     9.000

Sample Correlation Variance
     .0357

Sampling Error Variance
     .0088

Estimated Variance in Population Correlations
     .0270

**********   PUBLICATION BIAS DIAGNOSTIC INDICATORS   **********

Rosenthal Fail-Safe N
       915
```

Figure 1. SPSS output for the syntax file Meta_Basic_r.sps for Cartwright-Hatton *et al.*'s (2004) systematic review.

predicted from the transformed moderator effect (represented by β_1):

$$z_r = \beta_0 + C\beta_1 + e_i. \tag{17}$$

The within-study error variance is represented by e_i which will on average be zero with a variance of $1/(n_i - 3)$. To calculate the moderator effect, β_1, a generalized least squares (GLS) estimate is calculated. For the purposes of this tutorial, it is not necessary to know the mathematics behind the process (if you are interested then read Field, 2003b; Overton, 1998). The main thing to understand is that the moderator effect is coded using contrast weights that relate to the moderator effect (like contrast weights in the

analysis of variance). In the case of a moderator effect with two levels (e.g., whether the CBT used was group therapy or individual therapy), we could give one level codes of −1, and the other level codes of 1 (you should use 0.5 and −0.5 if you want the resulting beta to represent the actual difference between the effect of group and individual CBT). As such, when we run a moderator analysis using SPSS we have to define contrast codes that indicate which groups are to be compared.

A cautionary tale: The risk of confounded inference caused by unequal cell sizes

For theoretical and practical reasons, the primary studies in a meta-analysis tend to focus on some combinations of levels of variables more than on others. For example, white people aged around 20 are more commonly used as participants in primary studies than black people aged around 50. The occurrence of unequal cell sizes can introduce spurious correlations between otherwise independent variables. Consider a meta-analysis of 12 primary studies that investigated the difference between active and passive movement in spatial learning using the effect size measure *d*. Two moderator variables were identified as potentially useful for explaining differences among studies: (a) whether a reward was offered for good performance, and (b) whether the spatial environment was real or virtual. However, only 8 out of the 12 studies provided information about the levels of the moderator variables employed in their particular cases. Table 3 presents the original data set with full information about all 12 studies that was not available to the meta-analyst. The design of the original data set is balanced, because cell sizes are equal. Table 3 also displays a reduced data set of eight studies, which has an unbalanced design because cell sizes are unequal.

Table 3. Artefactual effect of reward moderator owing to unequal cell sizes

Original balanced data set			Reduced unbalanced data set		
Reward	Environment	*d*	Reward	Environment	*d*
1	1	.58			
1	1	.60	1	1	.60
1	1	.62			
1	−1	.18	1	−1	.18
1	−1	.20	1	−1	.20
1	−1	.22	1	−1	.22
−1	1	.58	−1	1	.58
−1	1	.60	−1	1	.60
−1	1	.62	−1	1	.62
−1	−1	.18			
−1	−1	.20	−1	−1	.20
−1	−1	.22			
	Mean reward	.40		Mean reward	.30
	Mean no reward	.40		Mean no reward	.50

Note. Reward levels are (1 = yes, −1 = no), and environment levels are (1 = real, −1 = virtual).

In the original balanced data set, the mean effect of a real environment is greater than that of a virtual one (.6 vs. .2). Second, there is no difference between the mean

effect when reward is present and when it is absent (.4 vs. .4). Third, the correlation between the reward and environment factors is $r = 0$, as would be expected in a balanced design.

However, the meta-analyst must work with the reduced unbalanced data set because key information about levels of moderator variables is missing. In the reduced data set, the correlation between the reward and environment factors equals $r = -.5$. Crucially, the non-zero correlation allows variance from the environment variable to be recruited by the reward variable. In other words, the non-zero correlation induces a spurious difference between the reward level mean effects (.3 vs. .5). The artefactual difference is generated because high-scoring real environments are underrepresented when reward is present (.60 vs. .18, .20, .22), while low-scoring virtual environments are underrepresented when reward is absent (.20 vs. .58, .60, .62).

Although the pool of potential moderator variables is often large for any given meta-analysis, not all primary studies provide information about the levels of such variables. Hence, in practice, only a few moderator variables may be suitable for analysis. In our example, suppose that too few studies provided information about the environment (real or virtual) for it to be suitable for use as a moderator variable. In that event, the spurious difference between the reward levels would remain, and would be liable to be misinterpreted as a genuine phenomenon. In our example, we have the benefit of knowing the full data set and therefore being able to see that the missing data were not random. However, a meta-analyst just has the reduced data set and has no way of knowing whether missing data are random or not. As such, missing data does not invalidate a meta-analysis *per se*, and does not mean that moderator analysis should not be done when data are missing in studies for certain levels of the moderator variable. However, it does mean that when studies at certain levels of the moderator variable are under- or unrepresented your interpretation should be restrained and the possibility of bias made evident to the reader.

Moderator analysis using SPSS

The macros we have supplied allow both continuous and categorical predictors (moderators) to be entered into the regression model that a researcher wishes to test. To spare the researcher the complexities of effect coding, the levels of a categorical predictor are coded using integers 1, 2, 3, . . . to denote membership of category levels 1, 2, 3, . . . of the predictor. The macros yield multiple regression output for both fixed- and random-effects meta-analytic models.

The Cartwright-Hatton *et al.* data set is too small to do a moderator analysis, so we will turn to our second example of Tenenbaum and Leaper (2002). Tenenbaum and Leaper were interested in whether the effect of parents' gender schemas on their children's gender-related cognitions was moderated by the gender of the experimenter. A SPSS file of their data can be downloaded from the website that accompanies this article (in this case Tenenbaum_&_Leaper_2002.sav). Load this data file into SPSS and you will see that the moderator variable (gender of the experimenter) is represented by a column labelled 'catmod' in which male researchers are coded with the number 2 and females with 1. In this example, we have just one column representing our sole categorical moderator variable, but we could add in other columns for additional moderator variables.

The main SPSS syntax file (in this case Meta_Mod_r.sps) is run using a much simpler launch file. From SPSS, open the syntax file Launch_Meta_Mod_r.sps. This file should

appear in a syntax window and comprises three lines:

```
cd "%HOMEDRIVE%%HOMEPATH%\Documents\Meta-Analysis".
insert file = "Meta_Mod_r.sps".
Moderator_r  r=r  n=n  conmods=()  catmods=(catmod).
```

The first line simply tells SPSS where to find your meta-analysis files.[7] The second line references the main syntax file for the moderator analysis. If this file is not in the '…\Documents\Meta-Analysis' directory then SPSS will return a 'file not found' error message. The final line is the most important because it contains parameters that need to be edited. The four parameters need to be set to the names of the corresponding variables in the active data file:

- r = the name of the variable containing the effect sizes. In the Tenenbaum data file, this variable is named 'r', so we would edit this to read r=r, if you had labelled this column 'correlation' then you would edit the text to say r=correlation, etc.
- n=the name of the sample size variable. In the Tenenbaum data file, this variable is named 'n', so we would edit this to read n=n, if you had labelled this column 'sample_size' in SPSS then you would edit the text to say n=sample_size, etc.
- conmods=names of variables in the data file that represent continuous moderator variables, e.g., conmods=(arousal accuracy). We have no continuous moderators in this example so we leave the inside of the brackets blank, e.g., conmods=( ).
- catmods=names of categorical moderator variables, e.g., catmods=(gender religion). In the Tenenbaum data file, we have one categorical predictor which we have labelled 'catmod' in the data file, hence, we edit the text to say catmods=(catmod).

On the top menu bar of this syntax file, click 'Run' and then 'All'. (The format of the launch file for *d* as an effect size is much the same except that there are two variables for sample size representing the two groups, n1 and n2, which need to be set to the corresponding variable names in SPSS, e.g., n1=n_group1 n2=n_group2.)

Figure 2 shows the resulting output. Tenenbaum and Leaper used a fixed-effects model, and the first part of the output replicates what they report (with the 95% confidence interval reported in parentheses throughout): there was an overall small to medium effect, $r = .16$ (.14, .18), and the gender of the researcher significantly moderated this effect, $\chi^2(1) = 23.72$, $p < .001$. The random-effects model tells a different story: there was still an overall small to medium effect, $r = .18$ (.13, .22); however, the gender of the researcher did not significantly moderate this effect, $\chi^2(1) = 1.18$, $p = .28$. Given the heterogeneity in the data, this random-effects analysis is probably the one that should have been done.

Estimating publication bias

Earlier on we mentioned that publication bias can exert a substantial influence on meta-analytic reviews. Various techniques have been developed to estimate the effect

[7] Remember in versions of Windows before Vista, this line should be edited to include 'My Documents' (i.e., 'cd "%HOMEDRIVE%%HOMEPATH%\My Documents\Meta-Analysis"') and that Mac users should replace this line with 'cd "Documents\Meta-Analysis"'.

```
**********   META-ANALYSIS OF CORRELATION COEFFICIENTS:  r   **********

Note:   The analysis has been conducted on Fisher-transformed correlations.
Note:   Statistics, e.g., b-coefficients, refer to Fisher-transformed correlations.
Note:   The overall mean has been back-transformed into the original r scale.

**********  FIXED EFFECTS REGRESSION ANALYSIS  **********

== MODEL WITHOUT PREDICTORS ==

OVERALL MEAN:  95% CONFIDENCE BOUNDS,  TEST DIFFERENCE FROM ZERO
      Mean     Lower     Upper         t         p         n
      .162      .142      .182    16.253      .000    48.000


ESTIMATED EFFECT SIZE VARIANCE  (OVERALL HETEROGENEITY):  Q STATISTIC
  Variance      Chi2        df         p
      .011   151.071    47.000      .000


== MODEL WITH PREDICTORS ==

PREDICTOR NUMBERING
   .000 = Constant
  1.000 = catmod

CONTINUOUS PREDICTORS:  B-COEFFICIENT, 95% CONFID. BOUNDS, STAND. ERROR, T-TEST
  Predictr   B-Coeff     Lower     Upper   Std Err        df         t         p
      .000      .166      .146      .187      .010    45.000    16.499      .000


CATEGORICAL PREDICTORS:  CHI-SQUARED TEST
  Predictr      Chi2        df         p
     1.000    23.719     1.000      .000


RESIDUAL VARIATION:   QE STATISTIC  (GOODNESS OF FIT)
      Chi2        df         p
   127.352    46.000      .000


**********  RANDOM EFFECTS REGRESSION ANALYSIS  **********

== MODEL WITHOUT PREDICTORS ==

OVERALL MEAN:  95% CONFIDENCE BOUNDS,  TEST DIFFERENCE FROM ZERO
      Mean     Lower     Upper         t         p         n
      .175      .133      .216     8.343      .000    48.000

 == MODEL WITH PREDICTORS ==

PREDICTOR NUMBERING
   .000 = Constant
  1.000 = catmod

CONTINUOUS PREDICTORS:  B-COEFFICIENT, 95% CONFID. BOUNDS, STAND. ERROR, T-TEST
  Predictr   B-Coeff     Lower     Upper   Std Err        df         t         p
      .000      .175      .135      .215      .020    45.000     8.737      .000


CATEGORICAL PREDICTORS:  CHI-SQUARED TEST
  Predictr      Chi2        df         p
     1.000     1.184     1.000      .277


RESIDUAL VARIATION:   QE STATISTIC  (GOODNESS OF FIT)
      Chi2        df         p
    55.783    46.000      .153
```

Figure 2. SPSS output for the moderator analysis on Tenenbaum and Leaper's (2002) data using the r_metareg macro.

of this bias, and to correct for it. We will focus on only a selection of these methods. The earliest and most commonly reported estimate of publication bias is Rosenthal's (1979) fail-safe *N*. This was an elegant and easily understood method for estimating the number of unpublished studies that would need to exist to turn a significant population effect size estimate into a non-significant one. To compute Rosenthal's fail-safe *N*, each

effect size is first converted into a z score and the sum of these scores is used in the following equation:

$$N_{\text{fs}} = \frac{\left(\sum_{i=1}^{k} z_i\right)^2}{2.706} - k. \tag{18}$$

Here, k is the number of studies in the meta-analysis and 2.706 is intrinsic to the equation. For Cartwright-Hatton *et al.*'s data, we get 915 from our SPSS basic analysis syntax (see Figure 1). In other words, there would need to be 915 unpublished studies not included in the meta-analysis to make the population effect size non-significant.

However, the fail-safe N has been criticized because of its dependence on significance testing. As testing the significance of the estimate of the population effect size is not recommended, other methods have been devised. For example, when using d as an effect size measure, Orwin (1983) suggests a variation on the fail-safe N that estimates the number of unpublished studies required to bring the average effect size down to a predetermined value. This predetermined value could be 0 (no effect at all), but could also be some other value that was meaningful within the specific research context: for example, how many unpublished studies there would need to be to reduce the population effect size estimate from .67 to a small, by Cohen's (1988) criterion, effect of .2. However, any fail-safe N method addresses the wrong question: it is usually more interesting to know the bias in the data one has and to correct for it than to know how many studies would be needed to reverse a conclusion (see Vevea & Woods, 2005).

A simple and effective graphical technique for exploring potential publication bias is the funnel plot (Light & Pillemer, 1984). A funnel plot displays effect sizes plotted against the sample size, standard error, conditional variance, or some other measure of the precision of the estimate. An unbiased sample would ideally show a cloud of data points that is symmetric around the population effect size and has the shape of a funnel.

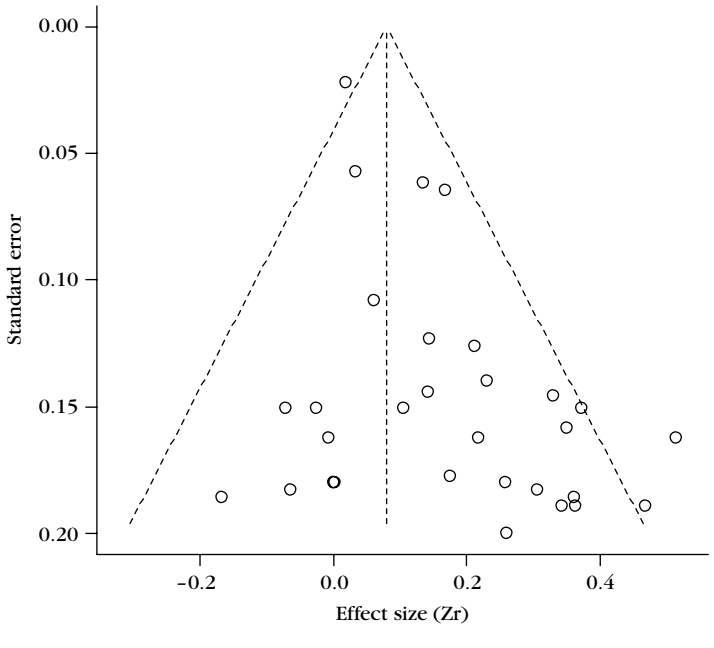

Figure 3. Example of a funnel plot showing little publication bias. The vertical line is the population effect size estimate and the diagonal lines the 95% confidence interval.

This funnel shape reflects the greater variability in effect sizes from studies with small sample sizes/less precision. A sample with publication bias will lack symmetry because studies based on small samples that showed small effects will be less likely to be published than studies based on the same-size samples but that showed larger effects (Macaskill, Walter, & Irwig, 2001). Figure 3 presents an example of a funnel plot showing approximate symmetry around the population effect size estimate. When you run the SPSS syntax for the basic meta-analysis, funnel plots are produced; however, the *y*-axis is scaled the opposite way to normal conventions. For this reason, we advise that you use these plots only as a quick way to look for publication bias, and use our publication bias scripts in R to produce funnel plots for presentation purposes (see below). Funnel plots should really be used only as a first step before further analysis because there are factors that can cause asymmetry other than publication bias. Some examples are true heterogeneity of effect sizes (in intervention studies this can happen because the intervention is more intensely delivered in smaller more personalized studies), English language bias (studies with smaller effects are often found in non-English language journals and get overlooked in the literature search) and data irregularities including fraud and poor study design (Egger, Smith, Schneider, & Minder, 1997).

Attempts have been made to quantify the relationship between effect size and its associated variance. An easy method to understand and implement is Begg and Mazumdar's (1994) rank correlation test for publication bias. This test is Kendall's tau applied between a standardized form of the effect size and its associated variance. The resulting statistic (and its significance) quantifies the association between the effect size and the sample size: publication bias is shown by a strong/significant correlation. This test has good power for large meta-analyses but can lack power for smaller meta-analyses, for which a non-significant correlation should not be seen as evidence of no publication bias (Begg & Mazumdar, 1994). This statistic is produced by the basic meta-analysis syntax file that we ran earlier. In your SPSS output, you should find that Begg and Mazumdar's rank correlation for the Cartwright-Hatton *et al.* data is highly significant, $\tau(N = 10) = -.51$, $p < .05$, indicating significant publication bias. Similar techniques are available based on testing the slope of a regression line fitted to the funnel plot (Macaskill *et al.*, 2001).

Funnel plots and the associated measures of the relationship between effect sizes and their associated variances offer no means to correct for any bias detected. Two main methods have been devised for making such corrections. Trim and fill (Duval & Tweedie, 2000) is a method in which a biased funnel plot is truncated and the number (k) of missing studies from the truncated part is estimated. Next, k artificial studies are added to the negative side of the funnel plot (and therefore have small effect sizes) so that in effect the study now contains k studies with effect sizes as small in magnitude as the k largest effect sizes. A new estimate of the population effect size is then calculated including these artificially small effect sizes. This is a useful technique but, as Vevea and Woods (2005) point out, it relies on the strict assumption that all of the 'missing' studies are those with the smallest effect sizes; as such it can lead to overcorrection. More sophisticated correction methods have been devised based on weight function models of publication bias. These methods use weights to model the process through which the likelihood of a study being published varies (usually based on a criterion such as the significance of a study). The methods are quite technical and have typically been effective only when meta-analyses contain relatively large numbers of studies ($k > 100$). Vevea and Woods' (2005) recent method, however, can be applied to smaller meta-analyses and has relatively more flexibility for the meta-analyst to specify the likely

conditions of publication bias in their particular research scenario. Vevea and Woods specify four typical weight functions which they label 'moderate one-tailed selection', 'severe one-tailed selection' 'moderate two-tailed selection', and 'severe two-tailed selection'; however, they recommend adapting the weight functions based on what the funnel plot reveals (see Vevea & Woods, 2005).

Estimating and correcting for publication bias using a computer

We have already calculated the fail-safe *N*, Begg and Mazumdar's rank correlation and some crude funnel plots in our basic analysis. However, for the more sophisticated meta-analyst, we recommend producing funnel plots with confidence intervals superimposed, and correcting population effect size estimates using Vevea and Woods's methods (above). Vevea and Woods (2005) have produced code for implementing their sensitivity analysis in S-PLUS, and this code will also run in R.[8] We have produced script files for R that feed data saved from the initial SPSS meta-analysis into Vevea and Woods's S-PLUS code, and use the package 'meta' to produce funnel plots too (you can also use Mix or Review Manager to produce funnel plots).

To do this part of the analysis, you will need to download R, if you do not already have it, and install it.

You can now run a publication bias analysis on the Cartwright-Hatton *et al.* data. To do this, go to the 'File' menu in R and select 'Open document…'. Find your meta-analysis directory (remember that you created this folder earlier), and select the file Pub_Bias_r.R (remember that if *d* is your effect size then you must select the file Pub_Bias_d.R). This document will open in a new window within R. In this new window, simply click with the right mouse button and select 'Select all' and then click with the right mouse button again and select 'Run line or selection' (this process can be done more quickly by using the keyboard shortcut of Ctrl + A followed by Ctrl + R or cmd + A followed by cmd + enter on a Mac).

The resulting funnel plot (Figure 4) shows the effect size plotted against the standard error for each study, and a reference line representing the 95% confidence interval. If the data were unbiased, this plot would be funnel shaped around the dotted line and symmetrical. The resulting plot is clearly not symmetrical (and shows one effect size that appears to be very discrepant from the rest) or funnel shaped and shows clear evidence of bias.

Figure 5 shows the output for Vevea and Woods' (2005) sensitivity analysis for both fixed- and random-effects models. We will interpret only the random-effects model. The unadjusted population effect size estimate is first given (with its variance component) and also the value when this estimate is converted back into *r*. These values correspond approximately to the values that we have already calculated from our SPSS analysis. However, the adjusted parameter estimates provide the values of the population effect size estimate corrected for four different selection bias models outlined by Vevea and Woods (2005). The four different selection bias models represent a range of situations differing in the extent and form of the selection bias. As such, they are a reasonable starting-point in the absence of any better information

[8] Due to slight technical differences between S-PLUS and R, the code provided in Vevea and Woods (2005) will not run in R; however, the amended code for R cited here will run on both R and S-PLUS (thanks to Jack Vevea for these amendments). We have made some minor changes to the code to make it slightly more straightforward to use, and so that the results when using *r* are back-transformed from *z* to *r*.

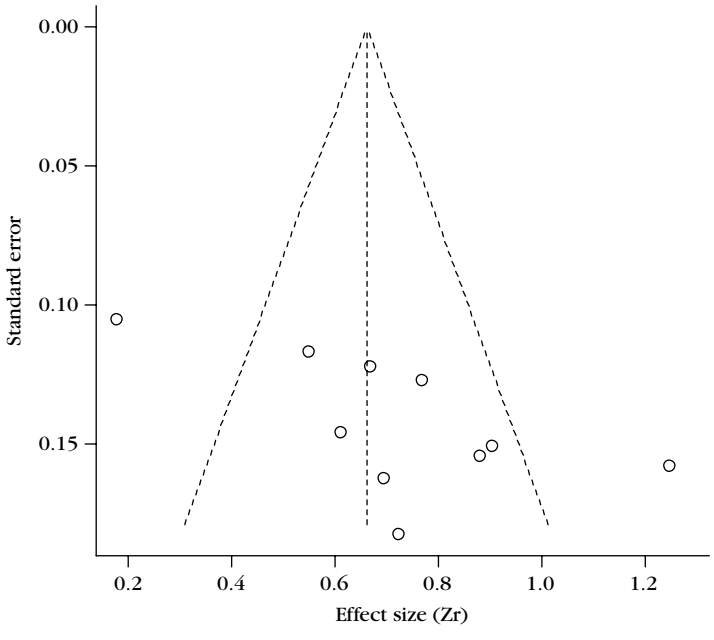

Figure 4. Funnel plot from R of effect sizes from Cartwright-Hatton *et al.* (2004).

```
**********  SENSITIVITY OF EFFECT-SIZE ESTIMATES TO PUBLICATION BIAS  **********


EFFECT-SIZE PARAMETER:   Correlation

EFFECT-SIZE ESTIMATOR:   r


FIXED-EFFECTS Publication Bias Model: Vevea & Woods (2005), Psychological Methods

Unadjusted Parameter Estimate
                                    zr         r
Parameter Estimates              0.66177441 0.5795431
Standard errors                  0.04331481        NA

Adjusted Parameter Estimate
                                   zr         r
Moderate One-Tailed Selection     0.6613350 0.5792511
Severe One-Tailed Selection       0.6608782 0.5789475
Moderate Two-Tailed Selection     0.6613340 0.5792505
Severe Two-Tailed Selection       0.6608738 0.5789446


RANDOM-EFFECTS Publication Bias Model: Vevea & Woods (2005), Psychological Methods
In this model v estimates population effect-size variance

Unadjusted Parameter Estimates
                                    zr          v        r
Parameter Estimates              0.71004644 0.05297219 0.610706
Standard errors                  0.08581562 0.03137239       NA

Adjusted Parameter Estimates
                                   zr          v         r
Moderate One-Tailed Selection     0.7027116 0.05282353 0.6060861
Severe One-Tailed Selection       0.6954048 0.05277298 0.6014429
Moderate Two-Tailed Selection     0.7034652 0.05283420 0.6065627
Severe Two-Tailed Selection       0.6960661 0.05277349 0.6018648
```

Figure 5. Output from R for Vevea and Woods' sensitivity analysis.

about the selection bias model most appropriate for your data (based on, for example, the funnel plot). However, Vevea and Woods (2005) recommend applying a greater variety of selection models, or applying selection models specifically tailored to the data within the particular meta-analysis.[9] The important thing in terms of interpretation is how the population effect size estimate changes under the different selection bias models. For the Cartwright-Hatton *et al.* data, the unadjusted population effect size (as *r*) was .61 as calculated above using the SPSS syntax. Under both moderate one- and two-tailed selection bias, the population effect size estimate is unchanged to two decimal places (see the column labelled *r* in Figure 5, for the random-effects model). Even applying a severe selection bias model, the population effect size drops only to .60. As such, we can be confident that the strong effect of CBT for childhood and adolescent anxiety disorders is not compromised even when applying corrections for severe selection bias.

Step 6: Write it up

Rosenthal (1995) wrote an excellent article on best practice in reporting meta-analytic reviews. Based largely on his advice, we recommend the following. First, you should always be clear about your search and inclusion criteria, which effect size measure you are using (and any issues you had in computing these), which meta-analytic technique you are applying to the data and why (especially whether you are applying a fixed- or random-effects method). Rosenthal recommends stem-and-leaf plots of the computed effect sizes because this is a concise way to summarize the effect sizes that have been included in your analyses. If you have carried out a moderator analysis, then you might also provide stem-and-leaf plots for subgroups of the analysis (e.g., see Brewin *et al.*, 2007). Other plots that should be considered are forest plots and a bean plot. You should always report statistics relating to the variability of effect sizes (these should include the actual estimate of variability as well as statistical tests of variability), and obviously the estimate of the population effect size and its associated confidence interval (or credibility interval). You should, as a matter of habit, also report information on publication bias, and preferably a variety of analyses (for example, the fail-safe *N*, a funnel plot, Begg and Mazumdar's rank correlation, and Vevea and Woods's sensitivity analysis).

Summary

This article has tried to offer a comprehensive overview of how to conduct a meta-analytic review including new files for an easy implementation of the basic analysis in

[9] These analyses can be achieved by editing the R file that runs the analysis. Opening this file in a text editor reveals the code that controls the weight functions:

```
# Enter fixed weight function
w1 <- matrix(c(1.0,.99,.95,.90,.80,.75,.65,.60,.55,.50,.50,.50,.50,.50), ncol = 1)
w2 <- matrix(c(1.0,.99,.90,.75,.60,.50,.40,.35,.30,.25,.10,.10,.10,.10), ncol = 1)
w3 <- matrix(c(1.0,.99,.95,.90,.80,.75,.60,.60,.75,.80,.90,.95,.99,1.0), ncol = 1)
w4 <- matrix(c(1.0,.99,.90,.75,.60,.50,.25,.25,.50,.60,.75,.90,.99,1.0), ncol = 1)
```

These four vectors contain the weights from Table 1 of Vevea and Woods (2005), and different selection bias models can be tested by changing the values of the weights in these vectors (see Vevea & Woods, 2005, for more details).

SPSS and R. To sum up, the analysis begins by collecting articles addressing the research question that you are interested in. This will include e-mailing people in the field for unpublished studies, electronic searches, searches of conference abstracts, and so on. Once the articles are selected, inclusion criteria need to be devised that reflect the concerns pertinent to the particular research question (which might include the type of control group used, clarity of diagnosis, the measures used, or other factors that ensure a minimum level of research quality). The included articles are then scrutinized for statistical details from which effect sizes can be calculated; the same effect size metric should be used for all studies (see the aforementioned electronic resources for computing these effect sizes). Next, decide on the type of analysis appropriate for your particular situation (fixed vs. random effects, Hedges' method or Hunter and Schmidt's, etc.) and then to apply this method (possibly using the SPSS resources produced to supplement this article). An important part of the analysis is to describe the effect of publication bias and to re-estimate the population effect under various publication bias models using the Vevea and Woods (2005) model. Finally, the results need to be written up such that the reader has clear information about the distribution of effect sizes (e.g., a stem-and-leaf plot), the effect size variability, the estimate of the population effect and its 95% confidence interval, the extent of publication bias (e.g., funnel plots, the rank correlation of the fail-safe *N*), and the influence of publication bias (Vevea and Woods's adjusted estimates).

Acknowledgements

Thanks to Jack Vevea for amending his S-PLUS code from Vevea and Woods (2005) so that it would run using R, and for responding to numerous e-mails from me about his weight function model of publication bias.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617. doi:10.1348/000712608X377117

Barrick, M. R., & Mount, M. K. (1991). The Big 5 personality dimensions and job-performance – a meta-analysis. *Personnel Psychology*, *44*(1), 1–26. doi:10.1111/j.1744-6570.1991.tb00688.x

Bax, L., Yu, L. M., Ikeda, N., Tsuruta, H., & Moons, K. G. M. (2006). Development and validation of MIX: Comprehensive free software for meta-analysis of causal research data. *BMC Medical Research Methodology*, *6*(50).

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101. doi:10.2307/2533446

Belanger, H. G., & Vanderploeg, R. D. (2005). The neuropsychological impact of sports-related concussion: A meta-analysis. *Journal of the International Neuropsychological Society*, *11*(4), 345–357.

Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*(4), 406–418. doi:10.1037/1082-989X.8.4.406

Brewin, C. R., Kleiner, J. S., Vasterling, J. J., & Field, A. P. (2007). Memory for emotionally neutral information in posttraumatic stress disorder: A meta-analytic investigation. *Journal of Abnormal Psychology*, *116*(3), 448–463. doi:10.1037/0021-843X.116.3.448

Cartwright-Hatton, S., Roberts, C., Chitsabesan, P., Fothergill, C., & Harrington, R. (2004). Systematic review of the efficacy of cognitive behaviour therapies for childhood and adolescent anxiety disorders. *British Journal of Clinical Psychology*, *43*, 421–436. doi:10.1348/0144665042388928

Clark-Carter, D. (2003). Effect size: The missing piece in the jigsaw. *Psychologist*, *16*(12), 636–638.

Cochrane Collaboration (2008). *Review Manager (RevMan) for Windows: Version 5.0*. Copenhagen: The Nordic Cochrane Centre.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037/0033-2909.112.1.155

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology*, *17*, 136–137. doi:10.1037/0735-7028.17.2.136

DeCoster, J. (1998). *Microsoft Excel spreadsheets: Meta-analysis*. Retrieved from http://www.stat-help.com/spreadsheets.html

Dickersin, K., Min, Y.-I., & Meinert, C. L. (1992). Factors influencing publication of research results: Follow-up of applications submitted to two institutional review boards. *Journal of the American Medical Association*, *267*, 374–378. doi:10.1001/jama.267.3.374

Douglass, A. B., & Steblay, N. (2006). Memory distortion in eyewitnesses: A meta-analysis of the post-identification feedback effect. *Applied Cognitive Psychology*, *20*(7), 859–869. doi:10.1002/acp.1237

Duval, S. J., & 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*(449), 89–98. doi:10.2307/2669529

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629–634.

Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin*, *132*(1), 33–72. doi:10.1037/0033-2909.132.1.33

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. doi:10.1037/1082-989X.6.2.161

Field, A. P. (2003a). Can meta-analysis be trusted? *Psychologist*, *16*(12), 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. doi:10.1207/S15328031US0202_02

Field, A. P. (2005a). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

Field, A. P. (2005b). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*(4), 444–467. doi:10.1037/1082-989X.10.4.444

Field, A. P. (2005c). 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., & Gillett, R. (2009). *How to do a meta-analysis*. Retrieved from http://www.statisticshell.com/meta_analysis/How_To_Do_Meta-Analysis.html

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675–701. doi:10.2307/2279372

Gillett, R. (2003). The metric comparability of meta-analytic effect-size estimators from factorial designs. *Psychological Methods*, *8*, 419–433. doi:10.1037/1082-989X.8.4.419

Greenwald, A. G. (1975). Consequences of prejudice against null hypothesis. *Psychological Bulletin*, *82*(1), 1–19. doi:10.1037/h0076157

Hafdahl, A. R. (2009). Improved Fisher *z* estimators for univariate random-effects meta-analysis of correlations. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 233–261. doi:10.1348/000711008X281633

Hafdahl, A. R. (2010). Random-effects meta-analysis of correlations: Monte Carlo evaluation of mean estimators. *British Journal of Mathematical and Statistical Psychology*, *63*, 227–254. doi:10.1348/000711009X431914

Hafdahl, A. R., & Williams, M. A. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological Methods*, *14*(1), 24–42. doi:10.1037/a0014697

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random-effects methods of meta-analysis. *Journal of Applied Psychology*, *87*(2), 377–389. doi:10.1037/0021-9010.87.2.377

Hedges, L. V. (1984). Estimation of effect size under non-random sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*, 61–85. doi:10.2307/1164832

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics*, *17*(4), 279–296. doi:10.2307/1165125

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217. doi:10.1037/1082-989X.6.3.203

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. doi:10.1037/1082-989X.3.4.486

Hox, J. J. (2002). *Multilevel analysis, techniques and applications*. Mahwah, NJ: Erlbaum.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8*(4), 275–292. doi:10.1111/1468-2389.00156

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

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594–612. doi:10.1037/0021-9010.91.3.594

Kelley, K., Bond, R., & Abraham, C. (2001). Effective approaches to persuading pregnant women to quit smoking: A meta-analysis of intervention evaluation studies. *British Journal of Health Psychology*, *6*, 207–228. doi:10.1348/135910701169160

Kontopantelis, E., & Reeves, D. (2009). MetaEasy: A meta-analysis add-in for Microsoft Excel. *Journal of Statistical Software*, *30*(7), 1–25.

Lavesque, R. (2001). *Syntax: Meta-analysis*. Retrieved from http://www.spsstools.net/

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistician*, *55*(3), 187–193. doi:10.1198/000313001317098149

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine*, *20*(4), 641–654. doi:10.1002/sim.698

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods*, *11*(4), 386–401. doi:10.1037/1082-989X.11.4.386

McLeod, B. D., & Weisz, J. R. (2004). Using dissertations to examine potential bias in child and adolescent clinical trials. *Journal of Consulting and Clinical Psychology*, *72*(2), 235–251. doi:10.1037/0022-006X.72.2.235

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*(2), 622–646. doi:10.1111/j.1467-8624.2007.01018.x

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

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, *8*, 157–159. doi:10.2307/1164923

Osburn, H. G., & Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology*, *77*(2), 115–122. doi:10.1037/0021-9010.77.2.115

Oswald, F. L., & McCloy, R. A. (2003). Meta-analysis and the art of the average. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 311–338). Mahwah, NJ: Erlbaum.

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*(3), 354–379. doi:10.1037/1082-989X.3.3.354

R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. doi:10.1037/0033-2909.86.3.638

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*(2), 183–192. doi:10.1037/0033-2909.118.2.183

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82. doi:10.1146/annurev.psych.52.1.59

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects – The first 345 studies. *Behavioral and Brain Sciences*, *1*(3), 377–386.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448–467. doi:10.1037/1082-989X.8.4.448

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97–128. doi:10.1348/000711007X255327

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge, MA: Hogrefe and Huber.

Schwarzer, G. (2005). *Meta*. Retrieved from http://www.stats.bris.ac.uk/R/

Shadish, W. R. (1992). Do family and marital psychotherapies change what people do? A meta-analysis of behavioral outcomes. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, *et al.* (Eds.), *Meta-analysis for explanation: A casebook* (pp. 129–208). New York: Russell Sage Foundation.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. doi:10.2307/2282137

Takkouche, B., Cadarso-Suárez, C., & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology*, *150*(2), 206–215.

Tenenbaum, H. R., & Leaper, C. (2002). Are parents' gender schemas related to their children's gender-related cognitions? A meta-analysis. *Developmental Psychology*, *38*(4), 615–630. doi:10.1037/0012-1649.38.4.615

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using *a priori* weight functions. *Psychological Methods*, *10*(4), 428–443. doi:10.1037/1082-989X.10.4.428

Whitener, E. M. (1990). Confusion of confidence-intervals and credibility intervals in metaanalysis. *Journal of Applied Psychology*, *75*(3), 315–321.

Wilson, D. B. (2004). *A spreadsheet for calculating standardized mean difference type effect sizes*. Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

Received 1 August 2007; revised version received 3 July 2009