Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2011, **2**, 1–10 doi: 10.1111/j.2041-210X.2010.00056.x

Getting started with meta-analysis

Freya Harrison*

Department of Zoology, University of Oxford, Oxford, UK

Summary

1. Meta-analysis is a powerful and informative tool for basic and applied research. It provides a statistical framework for synthesizing and comparing the results of studies which have all tested a particular hypothesis. Meta-analysis has the potential to be particularly useful for ecologists and evolutionary biologists, as individual experiments often rely on small sample sizes due to the constraints of time and manpower, and therefore have low statistical power.

2. The rewards of conducting a meta-analysis can be significant. It can be the basis of a systematic review of a topic that provides a powerful exploration of key hypotheses or theoretical assumptions, thereby influencing the future development of a field of research. Alternatively, for the applied scientist, it can provide robust answers to questions of ecological, medical or economic significance. However, planning and conducting a meta-analysis can be a daunting prospect and the analysis itself is invariably demanding and labour intensive. Errors or omissions made at the planning stage can create weeks of extra work.

3. While a range of useful resources is available to help the budding meta-analyst on his or her way, much of the key information and explanation is spread across different articles and textbooks. In order to help the reader use the available information as efficiently as possible (and so avoid making time-consuming errors) this article aims to provide a 'road map' to the existing literature. It provides a brief guide to planning, organizing and implementing a meta-analysis which focuses more on logic and implementation than on maths; it is intended to be a first port of call for those interested in the topic and should be used in conjunction with the more detailed books and articles referenced. In the main, references are cited and discussed with an emphasis on useful reading order rather than a chronological history of meta-analysis and its uses.

4. No prior knowledge of meta-analysis is assumed in the current article, though it is assumed that the reader is familiar with ANOVA and regression-type statistical models.

Key-words: effect size, meta-analysis, null hypothesis significance testing, power, *P*-values, sample size, statistics, systematic review

Meta-analysis makes me very happy – Jacob Cohen, psychologist and statistician

Introduction: the foundations of meta-analysis

A literature review for any given topic is likely to turn up a long list of studies, with varying degrees of consistency in experimental methodology, study species and analytical approach. Often these studies have led to very different conclusions. For example, theoreticians working on the evolution of biparental care have predicted that it is only an evolutionarily stable strategy if individuals respond to a decrease in parental care effort by their mate with an increase of smaller magnitude in their own care effort. Over the last 25 years, many behavioural ecologists have performed experiments to test whether partial compensation is indeed observed if one member of a breeding pair is removed or handicapped to reduce its care input. These studies have been carried out on birds, rodents and insects and have reported every possible response to experimental manipulation, from desertion to over-compensation for the lost care effort (reviewed in Harrison *et al.* 2009). Given the variability in how these studies were conducted and the often small individual sample sizes, it is almost impossible to decide if the literature as a whole supports the partial compensation hypothesis simply by reading and contrasting studies. However, **meta-analysis** provides a formal statistical framework with which we can rigorously combine and compare the results of these experiments. In this article, I will outline the logic of meta-analysis

*Correspondence author. E-mail: freya.andersdottir@gmail.com
Correspondence site: http://www.respond2articles.com/MEE

and provide a brief guide to planning, organizing and implementing a meta-analysis. The article is intended to serve as a 'road map' to the numerous detailed resources which are available, providing an introduction which focuses more on logic and implementation than on mathematics. A glossary of key terms (marked in bold in the main text) is provided in Box 1 and key references are listed at the end of each section.

Meta-analysis gives us quantitative tools to do two things. First, if a number of attempts have been made to measure the effect of one variable on another, then meta-analysis provides a method to calculate the mean effect of the independent variable, across all attempts. Usually, the independent variable represents some form of experimental manipulation (treated vs. control groups, or a continuous variable representing treatment level). To illustrate this, Fernandez-Duque & Valeggia (1994) combined the results of five studies of the effect of selective logging on bird populations. This revealed a detrimental effect of selective logging on population density that was not immediately apparent from simply looking at the results of the individual studies. Secondly, meta-analysis allows us to measure the amount of experimentally-induced change in the dependent variable across studies and to attempt to explain this variability using defined moderator variables. Such variables could reflect phylogenetic, ecological or methodological differences between study groups. For example, in a classic meta-analysis of 20 studies, Côté & Sutherland (1997) calculated that, on average, predator removal resulted in an increase in post-breeding bird populations but not in breeding populations.

Meta-analysis achieves these goals by using **effect sizes**: these are statistics that provide a standardized, directional measure of the mean change in the dependent variable in each study. Effect sizes can incorporate considerations of sample size. Furthermore, when being combined in a meta-analysis, effect sizes can be weighted by the variance of the estimate, such that studies with lower variance (i.e. tighter estimated effect size) are given more weight in the data set. Because variance decreases as sample size increases, this generally means that effect sizes based on larger study populations are given greater weight. Tests which are analogous to analysis of variance (ANOVA) and weighted regression can then be applied to the population of effect sizes to identify dependent variables that explain a significant amount of variation between studies. For instance, in our meta-analysis, we found that the mean response to partner removal or handicapping was indeed partial compensation and that the sex of the responding partner and aspects of the experimental methodology explained some of the variation between individual studies (Harrison *et al*. 2009).

Key references. Stewart (2010) and Hillebrand (2008) provide neat introductions to the logic and power of meta-analysis from an ecological perspective.

Meta-analysis vs. vote counting

How many of us have heard the results of **null hypothesis significance tests** (NHST) being referred to as showing 'strong' or 'weak' effects of some variable, based on the size of the calculated *P*-value? It is a common fallacy to assume that the smaller the *P*-value, the stronger the observed relationship must be. However, the magnitude of an effect and its statistical significance are not intrinsically correlated: a small *P*-value does not necessarily mean that the effect of experimental treatment is large, or that the slope of a variable of interest on some covariate is steep. This is due in large part to the dependence of *P* on sample size: given a large enough sample size, the null hypothesis will almost always be rejected. *P*-values reflect a dichotomous question (is the observed pattern of data likely to be due to chance, or not?) not an open-ended one (how strong is the pattern in the data?). Cohen (1990) uses a rather wonderful example to demonstrate this point: he cites a study of 14 000 children that reported a significant link between height (measured in feet and adjusted for age and sex) and IQ. He then points out that if we take 'significant to mean a *P*-value of 0·001, then a correlation coefficient of at least 0·0278 – a very shallow slope indeed – would be found to be significant in a sample this large. The authors actually reported a rather larger correlation coefficient of 0·11, but the effect of height on IQ is still small: converting height to a metric measure, this means that a 30-point increase in IQ would be associated with an increase in height of over a metre.

The pros and cons of NHST and its alternatives have been discussed by other authors (Nakagawa & Cuthill 2007; Stephens, Buskirk & del Rio 2007) and are beyond the scope of this article: suffice it to say that *P*-values from NHST do not measure the magnitude of the effect of independent variables on dependent variables, are heavily influenced by sample size and are not generally comparable across studies. In other words, *P*-values are not effect sizes: two studies can have the same effect size but different *P*-values, or *vice versa*.

This means that *post hoc* analyses that rely on 'vote counting' of studies with significant and non-significant results are not very reliable. Vote counting has been a common method of determining support for a hypothesis, is often used in the introduction or discussion sections of empirical papers to provide an overview of the current state of a field or to justify new work, and is sometimes published under the erroneous title of meta-analysis. No quantitative estimate of the effect of interest is provided by vote counting. Furthermore, vote counting lacks statistical **power** for two reasons. First, the effect of sample size on *P*-value means that real but small effects may have been obscured by small sample size in the original studies. Secondly, simply counting votes with no attention to effect magnitude or sample size does nothing to rectify this lack of power. A formal meta-analysis ameliorates this problem. Not only are effect sizes more informative, they also represent continuous variables that can be combined and compared. A more subtle point is that NHST focuses on reducing the probability of **type I errors** (rejecting the null hypothesis when it is in fact true). **Type II errors** (failing to reject the null hypothesis when it is false) are not so tightly controlled for and this type of error can be of particular concern in fields such as conservation or medicine, where failing to detect an effect of, say, pesticide use on farm bird populations, could be more harmful than a type I

error. By definition, any method that increases the power of a test reduces the likelihood of making a type II error.

There are three commonly-used types of statistic that give reliable and comparable effect sizes for use in meta-analysis. All can be corrected for sample size and weighted by within-study variance. For studies that involve comparing a continuous response variable between control and experimental groups, the mean difference between the groups can be calculated. For studies that test for an effect of a continuous or ordinal categorical variable on a continuous response variable, the correlation coefficient can be used. Finally, for dichotomous response variables the risk ratio or odds ratio provides a measure of effect size. Once a population of effect sizes has been collected, it is possible to calculate the mean effect size and also a measure of the amount of between-study variation (**heterogeneity**, Q) in effect size.

We might therefore say that meta-analysis is more clearly needs driven and evidence based than simple vote counting. Box 2 provides a simple demonstration of how meta-analysis works. It should be noted that, like any statistical method, meta-analyses are only as good as the data used and can still suffer from both type I and type II errors: this is dealt with in more detail in the discussion of meta-analytic methods below.

Key references. Useful textbooks on meta-analysis include those by Borenstein *et al*. (2009), Cooper, Hedges & Valentine (2009) and Lipsey & Wilson (2001) and a forthcoming volume edited by Koricheva, Gurevitch & Mengerson (in press). The introductory chapters of these books all expand on the information discussed by Stewart (2010) and Hillebrand (2008). Koricheva *et al*.'s book is explicitly written for researchers in ecology and evolution; of the other three, Borenstein *et al*.'s book is probably the most accessible and Lipsey & Wilson's the most concise. Cohen (1990) discusses some of the issues discussed above in a short and illuminating article on using and interpreting statistical tests, while Nakagawa & Cuthill (2007) provide a detailed discussion of significance vs. effect size. For readers interested in early arguments against vote counting and the first applications of meta-analysis in evolution and ecology, see the seminal work of Gurevitch and Hedges (e.g. Hedges & Olkin 1980; Gurevitch *et al*. 1992; Gurevitch & Hedges 1993).

A 'to do' list for meta-analysis

At this point, it would be useful to outline the steps required to begin and carry out a meta-analysis.

1 Perform a thorough literature search for studies that address the hypothesis of interest, using defined keywords and search methodology. This includes searching for unpublished studies, for example by posting requests to professional newsletters or mailing lists.

2 Critically appraise the resulting studies and assess whether they should be included in the review. (Are they applicable? Is the study methodology valid? Do you have enough information to calculate an effect size?) Record the reasons for dropping any studies from your data set.

3 Choose an appropriate measure of effect size and calculate an effect size for each study that you wish to retain.

4 Enter these studies into a master data base which includes study identity, effect size(s), sample size(s) and information which codes each study for variables which you have reason to believe may affect the outcome of each study, or whose possible influence on effect size you wish to investigate (experimental design, taxonomic information on the study species, geographic location of study population, life-history variables of the species used etc). You should also record how you calculated the effect size(s) for each study (see below).

5 Use meta-analytic methods to summarize the cross-study support for the hypothesis of interest and to try to explain any variation in conclusions drawn by individual studies.

6 Assess the robustness and power of your analysis (likelihood of type I and type II errors).

Steps 1 and 2 reflect the fact that meta-analysis sits within the general methodological framework of the systematic review. Cooper, Hedges & Valentine (2009) argue that research synthesis based on systematic reviews can be viewed as a scientific discipline in its own right. As they rightly stress, a good systematic review follows exactly the same steps as an experiment: a problem is identified, a hypothesis or hypotheses formulated, a method for testing the hypothesis designed and, once applied, the results of this method are quantitatively analysed. The method itself can then be criticized. These steps allow the goals of systematic review in general, or meta-analysis in particular, to be met. It is difficult to argue that a review has usefully contributed to a field – whether it be by providing critical analysis of empirical results, highlighting key issues or addressing a conflict – if the review itself does not have a firm basis in a defined methodology for identifying, including and extracting information from the sources reviewed. A notable proponent of the systematic review approach in ecology, Stewart (e.g. Stewart, Coles & Pullin 2005; Pullin & Stewart 2006; Roberts, Stewart & Pullin 2006) has provided guidelines relevant to this field.

The keys to making meta-analysis as stress-free as possible are organization and planning. In particular, your list of potential moderator variables (step 4) should be clearly defined before you begin: it is far preferable to produce a data base which includes information that you later decide not to use, than to produce a data base that excludes a variable you later decide to explore, as the latter may require a second (or third, or fourth) trawl through your collection of studies to extract the necessary information. In the present article, I will now concentrate on the mechanics of carrying out a meta-analysis (steps 3, 5 and 6).

Key references. DeCoster (2004) and Lipsey & Wilson (2001, Chapters 2, 4 & 5) provide an excellent and comprehensive guide to literature searching and study coding. It may be helpful at this stage to read a 'model' meta-analysis: good examples for ecologists include Fernandez-Duque & Valeggia (1994), Côté & Sutherland (1997) or Cassey *et al*. (2005); many further examples can be found at the website of the Meta-analysis in Ecology and Evolution working group (http://www.nceas.ucsb.edu/meta/)

Choosing an appropriate effect size statistic

A meaningful measure of effect size will depend on the nature of the data being considered. Experimental and observational studies in ecology and evolution generally generate data that falls into one of three categories, and this determines which indices of effect size are appropriate. All of the indices of effect size outlined below have known sampling distributions (generally they are normalized) and this allows us to calculate their standard errors and construct confidence intervals.

1 *Continuous or ordinal data from two or more groups*. Data in this form are exemplified by treatment vs. control group comparisons and are generally presented and analysed using averages and measures of variance (mean and standard deviation, median and interquartile range, etc). In such cases, a measure of the difference between the group means is an appropriate effect size. The raw difference in means can be standardized by the pooled standard deviation; two commonly-used measures of standardized mean difference are Cohen's *d* and Hedges' *g*: these differ in the method used for calculating the pooled standard deviation but it should be noted that the *d* and *g* notation has been used interchangeably by some authors. Alternatively, when the data measure rates of change in independent groups (e.g. plant growth response in normal or elevated CO_2, body mass gain after supplementary feeding), the response ratio can be used. This measures the ratio of the mean change in one group to the mean change in the other. Like the standardized mean difference, it takes the standard deviations in the two groups into account. The response ratio is generally log-transformed prior to meta-analysis in order to linearize and normalize the raw ratios.

2 *Continuous or ordinal data which are a response to a continuous or ordinal independent variable*. Any data which are analysed using correlation or regression fall into this category. In this case, the correlation coefficient itself can be used as a measure of effect size. Generally, we are interested in a simple bivariate relationship (say, the effect of average daily rainfall on the laying date of great tits), and it may be that the studies in our data set also explore such a relationship. If a study reports the results of statistical tests which include other variables (such as average daily temperature during the breeding season), then we might use the partial correlation coefficient: the effect of rainfall on lay date if temperature is held fixed. (It may be that this is the only effect size we can calculate from the data available; if the published data allow us to calculate the simple bivariate correlation between rainfall and laying date, ignoring temperature, then we could argue that it would be better to use this as our effect size as it would be more directly comparable with the bivariate correlation coefficients retrieved from the other studies). Whichever type of correlation coefficient we use, Fisher's *z* transformation is generally applied in order to stabilize the variance among coefficients prior to meta-analysis.

3 *Binary response data*. Data that take the form of binary yes/no outcomes, such as nest success or survival to the next breeding season, are generally analysed using logistic regression or a chi-squared test. In this case, an appropriate measure of effect size is given by calculating the risk ratio or odds ratio. These types of effect size have been very rarely used in ecology and evolution, though they are common in medical research.

In some cases, more than one type of effect size can be meaningful. For instance, if an experiment involves applying some quantitatively-measurable level of treatment to the experimental group, then the experimental and control groups could meaningfully be compared using either standardized mean difference or a correlation coefficient. If different studies have applied different levels of the treatment to their experimental groups, the latter may be preferable. The 'best' measure of effect size must be judged based on its compatibility with the available raw data and its ease of interpretation.

Much of the labour in conducting a meta-analysis lies in calculating individual effect sizes for studies. All of the effect sizes mentioned above can be calculated from reported means, variances, SEs, correlation coefficients and frequencies. If these are not available then effect sizes can be calculated from reported *t*, *F* or Chi-squared statistics or from *P*-values. The exact formulae for calculating effect sizes from these data differ depending on the nature of the statistical tests and experimental designs from which they were taken (e.g. paired vs. unpaired *t*-test). This is explained rather thoroughly by DeCoster (2004) and Nakagawa & Cuthill (2007). In general, the more directly you can calculate an effect size – the less you have to infer by using test statistics and reconstructed statistical tables – the less error will be incorporated into your estimate of the effect size. It is also possible to convert between different measures of effect size.

While the actual mathematics of converting reported data into effect sizes is rendered fairly straightforward thanks to freely-available Microsoft Excel files and meta-analytic software packages, actually harvesting the necessary data from a library of published studies can be painstaking work. The number of studies that have to be discarded due to an inability to calculate a meaningful effect size based on the information available can be surprisingly high. Studies that do not give variance statistics, do not clearly state which statistical tests were used or even do not make sample size explicit all create headaches for the would-be meta-analyst.

Key references. The textbooks referenced above all outline various effect size calculations, as do Hillebrand (2008), DeCoster (2004) and Nakagawa & Cuthill (2007). Hedges, Gurevitch & Curtis (1999) provide an introduction to the use of response ratios for ecological data and Schielzeth (2010) presents a thoughtful and interesting perspective on the calculation and presentation of correlation coefficients. Lipsey & Wilson (2002) helpfully provide an Excel spreadsheet for calculating effect size, which complements the information provided in the Appendices of their textbook; a similar spreadsheet is provided by Thalheimer & Cook (2002). The software packages outlined in '*Partitioning and expaining heterogeneity*' can also calculate effect sizes from summarized data.

Analysing your data

There are three main steps to a meta-analysis: calculating the mean effect size, calculating a measure of between-study heterogeneity in effect size and partitioning this heterogeneity between moderator variables and error. The three papers suggested in 'A "to do" list for meta-analysis' as 'model' ecological meta-analyses all provide clear and helpful guides for carrying out your own meta-analysis. Côté & Sutherland (1997) and Fernandez-Duque & Valeggia (1994) use Cohen's *d* as their effect size while Cassey *et al*. (2005) use transformed correlation coefficients. Côté & Sutherland and Cassey *et al* give details of between-group heterogeneity calculations. In our meta-analysis of parental care (Harrison *et al*. 2009), my colleagues and I show and discuss results derived from Lipsey & Wilson's (2002) SPSS macros.

THE MEAN EFFECT SIZE

If we have effect sizes from N studies and we denote the effect size for the ith study ES_i and its variance $s^2_{ES_i}$, we can then calculate the mean effect size across studies. In order to give studies with lower variance more weight in the calculation of the mean, we multiply each individual effect size by the inverse of its variance $1/s^2(ES_i)$, henceforth denoted w_i for brevity. The weighted mean effect size ES and its standard error SE_{ES} are thus calculated as follows.

$$\overline{ES} = \frac{\sum(ES_i w_i)}{\sum w_i} \qquad \text{eqn 1}$$

$$SE_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}} \qquad \text{eqn 2}$$

Note that this is a fixed-effect calculation of ES as it does not take into account random sources of variance (see below for a discussion of fixed and random effects). Confidence intervals for the mean effect size can then be calculated as for the individual effect sizes, using critical values from a standard normal or t distribution. If the confidence interval does not include zero, then we conclude that on average, experimental manipulation has a significant effect on our response variable at the specified significance level. Once the mean effect size has been calculated, the next step is to determine whether the various individual effect sizes in our sample of studies are likely to estimate the same population mean effect size.

HETEROGENEITY ACROSS STUDIES: FIXED AND RANDOM EFFECTS

If our sample of effect sizes reflects a single population mean effect size, then a study's effect size will differ from the true population mean purely as a result of which individuals were studied and the distribution of effect sizes will be homogeneous. It is possible to test whether our set of effect sizes shows more heterogeneity than would be expected due to sampling error alone by calculating the Q (often called Q_{Total}) statistic. The significance of Q is tested using either a chi-squared distribution or randomization tests and the calculation of Q is covered in the textbooks cited in 'Meta-analysis vs. vote counting'.

Variance in effect size between studies and a significant Q-value may stem from one of two broad sources. On the one hand, systematic, identifiable sources of variation such as species used, sex of individuals etc. may cause heterogeneity in effect size. Such sources of variation can be identified and their impact on effect size can be quantified using statistical tests analogous to ANOVA or weighted regression – this is exactly how Côté & Sutherland (1997) showed that predator removal differentially affects breeding and non-breeding bird populations, as mentioned in 'Introduction: The foundations of meta-analysis'. An overview of how heterogeneity is quantified and partitioned between moderator variables is given in the next section of this article, but first it is necessary to discuss the second source of variation: essentially random or non-identifiable differences between studies.

Most experimental biologists will be familiar with fitting random factors in ANOVA-type analyses to produce a mixed model. For instance, in behavioural experiments with repeated measures on the same individuals, individual identity should be coded as a random variable if we wish to generalize our results from these specific individuals to the population as a whole: the true effect of the experiment is not assumed to be the same for all individuals and we are interested in the mean effect across individuals, rather than finding a single 'true' effect that holds for every member of the population. By analogy, individual studies in a meta-analysis could each have some idiosyncrasies that we cannot either reveal or include in our model (age or sex ratio of the population used, season during which observations were taken etc.) So just as we measure experimental effects on a sample of individual animals from a population, in meta-analysis we have measured a random sample of effect sizes from a population of true effect sizes. If this is the case, then there must be some component of the total variance in the data set that represents these random effects and we must incorporate this into our analysis if we wish to generalize our results beyond this specific set of studies.

Therefore, if we obtain a significant Q-value from the fixed-effects calculation of ES, we must proceed in one of three ways. First, we could continue to assume a fixed effects model but add the assumption that between-study variation is the result of systematic, identifiable moderator variables and then attempt to identify these. Secondly, we could assume that the variation between studies is random. In this case, we can use mathematical methods for estimating the random effects variance component, add this component to the variance statistic for each individual effect size and re-calculate the inverse variance weights in order to calculate a random-effects version of ES. Thirdly, we could assume that heterogeneity stems from both fixed and random sources. In this case, we can run meta-analytic models to test for the effects of fixed moderator variables, using inverse variance weights that have been adjusted for the estimated random effects component. This produces a mixed-effects model and a more robust test for the significance of moderator variables, because it does not exclude

the possibility that some of the heterogeneity remaining after a model has been fitted is due to systemic but unidentified sources of variation. The type I error rate is thus reduced, though this does come at the cost of a loss of some power in testing for the effects of moderator variables (Lipsey & Wilson, 2001). The software packages detailed in Section 5.4 allow the user to specify fixed- or random-effects models and/or calculate the random effects variance component.

It is worth noting that sometimes moderator variables can explain some of the variance in the data set even when there is no evidence for significant overall heterogeneity in a fixed-effects estimate of ES. Therefore, if there are good *a priori* reasons for supposing an effect of a moderator, it is arguably worth testing for this even when Q is not significant. A related issue is that of non-independence of effect sizes across studies; it may be advisable to control for studies being conducted by the same research group, for instance, by including this as a moderator variable. In a population of studies carried out on different species, non-independence arises from phylogenetic relatedness. The development of phylogenetic meta-analysis is gaining momentum but is beyond the scope of this article; Adams (2008) and Lajeunesse (2009) discuss methodologies for conducting such analyses.

Key references. Part 3 of Borenstein *et al*. (2009) and Chapters 6–8 of Lipsey & Wilson (2001) provide clear explanations of fixed vs. random effects models and provide macros (2002) to calculate the random effects variance component. Gurevitch & Hedges (1993) provide one of the first discussions of mixed-effects models in ecological meta-analysis.

PARTITIONING AND EXPLAINING HETEROGENEITY

If moderator variables cause effect sizes to differ on a study-by-study basis, the distribution of effect sizes will be heterogeneous. For instance, if our studies measure the effect of environmental disturbance on antagonistic interactions between individuals of a species, and if this effect is moderated by the population density, then a data set which includes studies on high- and low-density populations may show a bimodal distribution of effect sizes. In this case, the null hypothesis of homogeneity is rejected and a single mean is not the best way to describe the the population of true effect sizes. Sources of systematic variation between studies – the potential moderator variables identified when planning and coding the meta-analysis – must be investigated to see if they do indeed affect the response to experimental manipulation.

Variables that explain significant proportions of heterogeneity can be identified using statistical tests that are analogous to ANOVA or weighted regression, but which do not rely on the assumption of homogeneity of variance. If we have a categorical moderator variable, for example sex, then just as an ANOVA would partition total variance in a data set into variance due to the explanatory variables (between-sex variance) and variance due to error (within-sex variance), so heterogeneity can be split into between- and within-group components. If Q_{Between} is significantly greater than Q_{Within}, this indicates that sex explains a significant proportion of the total heterogeneity in effect sizes. The mean effect sizes within each sex can then be calculated. In the case of continuous or ordinal categorical moderator variables, an analogue to a weighted regression model can be used to determine whether fitting one or more of these explanatory variables explains a significant amount of heterogeneity (Q_{Model} vs. Q_{Residual}). In this case, for each variable treated as a covariate, an estimate of the slope of its effect and a corresponding P-value can also be calculated. Implementing these models is rendered fairly straightforward by the availability of specialist software such as MetaWin (Rosenberg, Adams & Gurevitch 2000) or Meta-analyst (Wallace *et al*. 2009) and by macros or extensions for common software packages (e.g. macros for SPSS: Lipsey & Wilson 2001, 2002; MIX for Microsoft Excel: Bax *et al*. 2008).

Criticizing meta-analysis

The robustness and utility of meta-analysis – and the reliability of any inferences drawn from it – are determined ultimately by the population of individual studies used. First, issues surrounding which studies can and should be included in a meta-analysis should be mentioned. Secondly, it would be useful to have some way of determining the likelihood of a significantly non-zero mean effect size being the result of a type I error and, conversely, the likelihood of a zero mean effect size being the result of a lack of statistical power rather than a reliable reflection of the true population mean effect size. The number and identity of studies used, as well as their individual sample sizes, will affect type I and II error rate in meta-analysis.

WHICH STUDIES CAN BE COMBINED IN A META-ANALYSIS?

Step 2 in the 'to do' list reflects the fact that including methodologically poor studies in the data set may add more noise than signal, clouding our ability to calculate a robust mean effect size or to identify important moderator variables. Defining and reporting the criteria by which studies were assessed for inclusion is therefore an essential part of the meta-analytic method. Furthermore, thought must be given as to whether the studies under consideration may sensibly be combined in a meta-analysis – do the effect sizes calculated from the population of studies all reflect the same thing? For instance, both feeding offspring and providing thermoregulation by means of brooding or incubating are types of parental care, but in our meta-analysis (Harrison *et al*. 2009) we considered these two types of care separately. The effect sizes for the two types of care were significantly different as defined by a Q test, but more fundamentally there is no reason to assume that these behaviours have the same cost : benefit ratios for parents: therefore, we felt that combining their effect sizes would be an example of 'comparing apples and oranges' – a criticism that has often been levelled at meta-analysis. This consideration is probably more pertinent to ecologists than to, say, medical researchers, as response variables and study designs vary more widely in our field.

It is also worth noting that individual studies may act as outliers in a meta-analytic data set, having a very large influence on the mean effect size. It is possible to identify such studies by means of a leave-out analysis: each of our *N* studies is dropped from the data set in turn and a set of estimates of the mean effect sizes from the *N-1* remaining studies is calculated. Software such as the aforementioned Meta-Analyst can perform an automated leave-out analysis and so flag highly influential studies. How to deal with such a study is then a matter for personal consideration; depending on the nature of the study (sample size, experimental protocol, apparent methodological quality), the meta-analyst must decide whether it is justifiable to leave it in the data set, or better to remove it. If it is retained, then it would be advisable to report the effect of dropping this study on the conclusions.

We must also consider potential sources of non-independence in the data set. Non-independence has already been mentioned in the context of moderator variables such as research group, and in the context of phylogenetic meta-analysis. However, non-independence can also result from more than one effect being measured on each individual or replicate in a study. For example, if we have data on reproductive success and survival in control and experimental groups, then including the whole population of effect sizes in a single analysis not only raises the issue of a potential 'apples and oranges' comparison, but also creates non-independence as a result of measures from the same individuals being correlated. In this scenario, arguably the best strategy is to conduct separate meta-analyses of effects on reproduction and survival. Non-independence also rears its head in another form if we test the same set of studies over and over for the effects of different moderator variables. This will compromise the reliability of our significance tests and increase the type I error rate.

PUBLICATION BIAS

The biggest potential source of type I error in meta-analysis is probably publication bias. A funnel plot of effect size vs. study size is one method of identifying publication bias in our set of studies: all thing being equal, we would expect that the effect sizes reported in a number of studies should be symmetrically distributed around the underlying true effect size, with more variation from this value in smaller studies than in larger ones. Asymmetry or gaps in the plot are suggestive of bias, most often due to studies which are smaller, non-significant or have an effect in the opposite direction from that expected having a lower chance of being published. A more thorough discussion of publication bias is provided by Sutton (2009). For the purposes of this article, suffice it to say that time spent uncovering unpublished data relevant to the hypothesis in question, as suggested in the 'to do' list above, is highly recommended.

Even if we discover and include some unpublished studies and produce a funnel plot with no glaring gaps, it would still be informative if we could work out the number of non-significant, unpublished studies that would have to exist, lying buried in file drawers and field notebooks, in order to make us suspect that our calculated mean effect size is the result of a type I error. This is termed the **failsafe sample size** and various simple, back-of-an-envelope methods have been suggested for calculating it, based on the number of studies included, their effect sizes and some benchmark minimal meaningful effect size. The larger the failsafe sample size, the more confident we can be about the representativeness of our data set and the robustness of any significant findings. However, Rosenberg (2005) makes the important point that suggested methods for calculating the failsafe sample size are overly simple and likely to be misleading, in the main because they do not take into account the weighting of individual studies in the meta-analytic data set – a curious omission, given that weighting is one of the key strengths of meta-analysis. He outlines a method for calculating the failsafe sample size which is arguably more explicitly 'meta-analytic' in its calculation.

The reader should therefore be aware that the utility of failsafe sample size calculations is still debated. Jennions, Møller & Hunt (2004) and Møller & Jennions (2001) provide an interesting discussion of publication bias and type I errors in meta-analysis. These authors stress the point that meta-analysis involves (or should involve) explicit consideration of publication bias and attempts to minimize its influence, and that this should primarily consist of seeking unpublished studies (as opposed to *post hoc* calculations). If I may venture a tentative opinion, I would suggest that a report of failsafe sample size is worth including in published meta-analyses, but it is no substitute for a thorough search for unpublished data and should be interpreted as only a rough reflection of the likely impact of any publication bias.

POWER

As discussed above, type II errors often concern us more than type I errors. If our mean effect size is not significantly different from zero, if no significant heterogeneity is found among studies, or if a moderator variable is concluded to have no effect on effect size, how can we start to decide if this is simply due to a lack of statistical power? Evaluating the power of meta-analytic calculations is rather more complex as it depends on both the number of studies used and their individual sample sizes, which are related to the within-study component of variance in effect size. Hedges & Pigott (2001, 2004) provide detailed guides to power calculations for meta-analysis. In the present article, I will limit the discussion of power to the observation that small studies which in themselves have low statistical power might add more noise than signal to a meta-analytic data set and thus reduce its power: the benefits of excluding studies with very small sample size should be seriously considered, and can be quantified by calculating the power of a meta-analytic data set that either includes or excludes such studies.

Key references. Most of the general references given in earlier sections also discuss criticisms and limitations of meta-analysis. A free program for carrying out the calculations described in Rosenberg (2005) is available from http://www.rosenberglab.net/software.php. Power calculations are discussed in Chapter 29 of Borenstein *et al.* (2009), in Lajeunesse's chapter

in the forthcoming book by Koricheva et al. and in more detail by Hedges & Pigott (2001, 2004) and Cafri & Kromrey (2008) have developed an SAS macro to calculate power using the methods described by Hedges & Pigott.

Closing remarks

Meta-analysis is a great tool for extracting as much information as possible from a set of empirical studies. The potential advantages of sharing and combining data in this way are, I hope, evident from the discussion in this article. Organizing and carrying out a meta-analysis is hard work, but the fruits of the meta-analyst's labour can be significant. In the best case scenario, meta-analysis allows us to perform a relatively powerful test of a specific hypothesis and to draw quantitative conclusions. A low-powered analysis based on a small number of studies can still provide useful insights (e.g. by revealing publication bias through a funnel plot). Finally, by revealing the magnitude of effect sizes associated with prior research, meta-analysis can suggest how future studies might best be designed to maximize their individual power.

Most journals now include in their instructions to authors a sentence to the effect that effect sizes should be given where appropriate, or that at least the necessary information required for rapidly calculating an effect size should be provided. The lack of this information is common, but will not necessarily be noticed by the authors, interested readers or peer reviewers. For example, when conducting our meta-analysis on parental care (Harrison *et al*. 2009), it was only on specifically attempting to calculate effect sizes that we noticed a small number of published articles where the sample sizes used were not clear. Double-checking that sample sizes are stated explicitly and that exact test statistics and *P*-values are stated should not add significantly to the burden of writing up a research article and will add value to the work by allowing its ready incorporation to a meta-analysis if required. On a more positive note, we received many rapid and positive responses from colleagues whom we contacted to ask for clarification, extra data or unpublished data. There is clearly a spirit of cooperation in ecology and evolution which can lead to the production of useful and interesting syntheses of key issues in the field.

Box 1: Glossary

Effect size: A standardized measure of the response of a dependent variable to change in an independent variable; often but not always a response to experimental manipulation. Effect sizes could be thought of as *P*-values that have been corrected for sample size and are the cornerstone of meta-analysis: they make statistical comparison of the results of different studies valid. Commonly-used effect size measurements are the standardized mean difference between control and experimental groups, correlation coefficients and response ratios.

Fail-safe sample size: If we calculate a mean effect size across studies and it is significantly different from zero, the failsafe sample size is the number of unpublished studies with an effect size of zero that would have to exist in order to make our significant result likely to be due to sampling error rather than any real effect of the experimental treatment. i.e. the bigger this value, the smaller the probability of a *type I error*. The utility of failsafe sample sizes is debated.

Heterogeneity: A measure of the among-study variance in effect size, denoted Q. Just as ANOVA-type statistical analyses partition variance between defined independent variables and error to perform significance tests, meta-analysis can partition heterogeneity between independent variables of interest and error.

Meta-analysis: A formal statistical framework for comparing the results of a number of empirical studies that have tested, or can be used to test, the same hypothesis. Meta-analysis allows us to calculate the mean response to experimental treatment across studies and to discover key variables that may explain any inconsistencies in the results of different studies.

Null hypothesis significance testing: 'Traditional' statistical tests are tools for deciding whether an observed relationship between two or more variables is likely to be caused simply by sampling error. A test statistic is calculated based on the variance components of the data set and compared with a known frequency distribution to determine how often the observed patterns in the data set would arise by chance, given random sampling from a homogeneous population.

Power: The ability of a given test using a given data set to reject the null hypothesis (at a specified significance level) if it is false. i.e. as power increases, the probability of making a *type II error* decreases.

Type I error: Rejecting the null hypothesis when it is true (see *Fail-safe sample size*).

Type II error: Failing to reject the null hypothesis when it is false (see *Power*).

Box 2: The power of meta-analysis

Imagine that a novel genetic polymorphism has been discovered in a species of mammal. It has been hypothesized that the 'mutant' genotype may affect female lifetime reproductive success (LRS) relative to the wild type. Twelve groups of researchers genotype a number of females and record their LRS. Each group studies equal numbers of wild-type and mutant females, with total sample sizes ranging from 18 to 32 animals. Six of the studies were carried out on one long-term study population in habitat A and six on a second in habitat B.

Unknown to the researchers, there is a habitat-dependent effect of genotype on female LRS. Across the whole species wild-type females produce on average 5·0 ± 2·0 offspring that survive to reproductive age. In habitat A, mutant females also produce 5·0 ± 2·0 offspring, but in habitat B mutant LRS is increased to 5·8 ± 2·0 offspring. The standardized mean difference in female LRS is, therefore, zero in habitat A and 0·4 in habitat B.

The results of the imaginary studies are given in the table below and are based on random sampling from normal distributions with the specified means and standard deviations (Table 1). For each study, LRS (mean and SD) is given for

Table 1. Results of 12 studies investigating effect of genotype on LRS

		Wild-type LRS			Mutant LRS			
Study	Habitat	Mean	SD	*N*	Mean	SD	*N*	*P* (two-tailed)
1	A	4·71	1·55	10	5·28	2·74	10	0·575
2	A	4·5	1·51	12	5	2·23	12	0·470
3	A	5·12	1·94	14	4·79	2·1	14	0·671
4	A	4·92	2·01	11	4·68	1·35	11	0·745
5	A	4·93	1·65	15	4·19	2·24	15	0·312
6	A	5·08	2·36	16	5·11	1·58	16	2·122
7	B	4·71	1·15	12	5·94	1·66	12	0·047
8	B	4·9	1·7	10	6·01	1·22	10	0·110
9	B	5·1	2·05	9	5·83	1·81	9	0·257
10	B	4·77	1·78	16	6·99	1·7	16	0·001
11	B	3·92	2·55	14	5·84	1·94	14	0·034
12	B	4·99	1·6	15	5·72	1·66	15	0·229

LRS, lifetime reproductive success.

each genotype, along with the sample sizes and the *P*-value resulting from a *t*-test. Based on *t*-tests, three studies reported a significant effect of the mutant allele on LRS.

Can we use meta-analytic techniques to combine these data and gain quantitative estimates for the size of the effect of genotype on LRS? Fig. 1 shows the calculated mean effect size (Cohen's *d*) for each study, with their 95% confidence intervals. The 95% confidence interval for the weighted mean effect size across all twelve studies is (0·06, 0·64), suggesting that the mutation does indeed increase LRS. Furthermore, if we treat habitat as a moderator variable, the genotype by environment interaction is revealed: the 95% confidence interval for the mean effect is (−0·35, 0·29) in habitat A and (0·40, 1·1) in habitat B. Thus the mean effect size is not significantly different from zero in habitat A, but positive in habitat B. Also, the confidence interval for habitat B (just) captures the 'true' effect size of 0·4.

This example should serve to demonstrate that meta-analysis is a powerful way of synthesizing data and effectively increasing sample size to provide a more robust test of a hypothesis. However, like all statistical methods, the results of meta-analysis should be interpreted in the light of various checks and balances which can inform us as to the likely reliability of our conclusions: this is discussed in the main text.

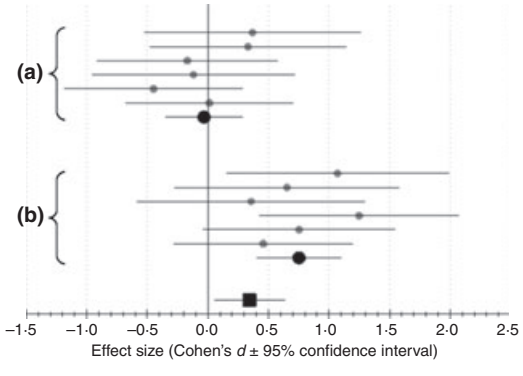

Fig. 1. Results of 12 studies investigating effect of genotype on lifetime reproductive success (LRS). Small grey circles show effect sizes from individual studies, large black circles mean effect sizes for the two habitats (A and B). The black square shows the overall mean effect size. All effect sizes are Cohen's *d* and 95% confidence interval. The fictitious data set was analysed using Meta-Analyst (Wallace *et al*. 2009).

Acknowledgements

I would like to thank my co-authors for my own first foray into meta-analysis, Zoltán Barta, Innes Cuthill and Tamás Székely, for their support. I would also like to thank Rob Freckleton, Michael Jennions and one anonymous reviewer for their helpful criticisms and suggestions, and finally Andy Morgan for proof-reading the manuscript.

References

Adams, D.C. (2008) Phylogenetic meta-analysis. *Evolution*, **62**, 567–572.

Bax, L., Yu, L.M., Ikeda, N., Tsuruta, H. & Moons, K.G.M. (2008) MIX: comprehensive free software for meta-analysis of causal research data. Version 1.7. http://mix-for-meta-analysis.info.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. & Rothstein, H.R. (2009) *Introduction to Meta-Analysis (Statistics in Practice)*. John Wiley & Sons, Chichester.

Cafri, G. & Kromrey, J.D. (2008) ST-159: A SAS® macro for statistical power calculations in meta-analysis. *SESUG 2008: The Proceedings of the South East SAS Users Group, St Pete Beach, FL, 2008*.

Cassey, P., Blackburn, T.M., Duncan, R.P. & Lockwood, J.L. (2005) Lessons from the establishment of exotic species: a meta-analytical case study using birds. *Journal of Animal Ecology*, **74**, 250–258.

Cohen, J. (1990) Things I have learned (so far). *American Psychologist*, **45**, 1304–1312.

Cooper, H.M., Hedges, L.V. & Valentine, J.C. (eds) (2009) *The Handbook of Research Synthesis and Meta-Analysis*, 2nd edn. Russell Sage Foundation, New York, NY.

Côté, I.M. & Sutherland, W.J. (1997) The effectiveness of removing predators to protect bird populations. *Conservation Biology*, **11**, 395–405.

DeCoster, J. (2004) Meta-analysis notes. Retrieved from http://www.stat-help.com/notes.html.

Fernandez-Duque, E. & Valeggia, C. (1994) Meta-analysis – a valuable tool in conservation research. *Conservation Biology*, **8**, 555–561.

Gurevitch, J. & Hedges, L.V. (1993) Meta-analysis: combining the results of independent experiments. *The Design and Analysis of Ecological Experiments* (eds S.M. Scheiner & J. Gurevitch), pp. 378–398. Chapman & Hall, London.

Gurevitch, J., Morrow, L.L., Wallace, A. & Walsh, J.S. (1992) A meta-analysis of field experiments on competition. *American Naturalist*, **140**, 539–572.

Harrison, F., Barta, Z., Cuthill, I. & Székely, T. (2009) How is sexual conflict over parental care resolved? A meta-analysis. *Journal of Evolutionary Biology*, **22**, 1800–1812.

Hedges, L.V., & Olkin, I. (1980) Vote counting methods in research synthesis. *Psychological Bulletin*, **88**, 359–369.

Hedges, L.V. & Pigott, T.D. (2001) The power of statistical tests in meta-analysis. *Psychological Methods*, **6**, 203–217.

Hedges, L.V. & Pigott, T.D. (2004) The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, **9**, 426–445.

Hedges, L.V., Gurevitch, J. & Curtis, P.S. (1999) The meta-analysis of response ratios in experimental ecology. *Ecology*, **80**, 1150–1156.

Hillebrand, H. (2008) Meta-analysis in ecology. *Encyclopedia of Life Sciences (ELS)*. John Wiley & Sons, Ltd, Chichester.

Jennions, M.D., Møller, A.P. and Hunt, J. (2004) Meta-analysis can "fail": reply to Kotiaho & Tomkins. *Oikos*, **104**, 191–193.

Koricheva, J. Gurevitch, J. & Mengerson, K. (eds) (in press) *Handbook of Meta-Analysis in Ecology and Evolution*. Princeton University Press, Princeton, NJ.

Lajeunesse, M.J. (2009) Meta-analysis and the comparative phylogenetic method. *American Naturalist*, **174**, 369–381.

Lipsey, M.W. & Wilson, D.B. (2001) *Practical Meta-analysis (Applied Social Research Methods Series Volume 49)*. SAGE Publications, Thousand Oaks, CA.

Lipsey, M.W. & Wilson, D.B. (2002) Effect size calculator and SPSS macros available from http://mason.gmu.edu/~dwilsonb/ma.html.

Møller, A.P. & Jennions, M.D. (2001) Testing and adjusting for publication bias. *Trends in Ecology and Evolution*, **16**, 580–586.

Nakagawa, S. & Cuthill, I.C. (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, **82**, 591–605.

Pullin, A.S., & Stewart, G.B. (2006) Guidelines for systematic review in conservation and environmental management. *Conservation Biology*, **20**, 1647–1656.

Roberts, P.D., Stewart, G.B., & Pullin, A.S. (2006) Are review articles a reliable source of evidence to support conservation and environmental management? A comparison with medicine. *Biological Conservation*, **132**, 409–423.

Rosenberg, M.S. (2005) The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, **59**, 464–468.

Rosenberg, M.S., Adams, D.C., & Gurevitch, J. (2000) MetaWin: statistical software for meta-analysis. Version 2.0. Software and manual available from http://www.metawinsoft.com.

Schielzeth, H. (2010) Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, **1**, 103–113.

Stephens, P.A., Buskirk, S.W. & del Rio, C.M. (2007) Inference in ecology and evolution. *Trends in Ecology & Evolution*, **22**, 192–197.

Stewart, G. (2010) Meta-analysis in applied ecology. *Biology Letters*, **6**, 78–81.

Stewart, G.B., Coles, C.F., & Pullin, A.S. (2005) Applying evidence-based practice in conservation management: lessons from the first systematic review and dissemination projects. *Biological Conservation*, **126**, 270–278.

Sutton, A.J. (2009) Publication bias. *The handbook of research synthesis and meta-analysis*, 2nd edn (eds H.M. Cooper, L.V. Hedges & J.C. Valentine), pp. 435–452. Russell Sage Foundation, New York, NY.

Thalheimer, W. & Cook, S. (2002) How to calculate effect sizes from published research articles: a simplified methodology. Retrieved from http://work-learning.com/effect_sizes.htm.

Wallace, B.C., Schmid, C.H., Lau, J. & Trikalinos, T.A. (2009) Meta-Analyst: software for meta-analysis of binary, continuous and diagnostic data. *BMC Medical Research Methodology*, **9**, 80.

Received 10 March 2010; accepted 28 June 2010
Handling Editor: Robert P. Freckleton