Meta-analysis in Ecology

JESSICA GUREVITCH, PETER S. CURTIS
AND MICHAEL H. JONES

I. SUMMARY

Meta-analysis is the statistical synthesis of the results of separate studies. It was adapted from other disciplines for use in ecology and evolutionary biology beginning in the early 1990s, and, at the turn of the century, has begun to have a substantial impact on the way data are summarized in these fields. We identify 119 studies concerned with meta-analysis in ecology and evolution, the earliest published in 1991 and the most recent in 2000. We introduce the statistical methods used in modern meta-analysis with references to the well-developed literature in the field. These formal, statistically defensible methods have been established to determine average treatment effects across studies when a common research question is being investigated, to establish confidence limits around the average effect size, and to test for consistency or lack of agreement in effect size as well as explanations for differences in the magnitude of the effect among studies. Problems with popular but statistically flawed methods for the quantitative summary of research results have been pointed out, and their use is diminishing. We discuss a number of challenges and threats to the validity of meta-analysis in ecology and evolution. In particular, we examine how difficulties resulting from missing data, publication bias, data quality and data exclusion, non-independence among observations, and the combination of dissimilar data sets may affect the perceived utility of meta-analysis in these fields and the soundness of conclusions drawn from its application. We highlight particular applications of meta-analysis in ecology and evolution, discuss several controversies surrounding individual meta-analyses, and outline some of the practical issues involved in carrying out a meta-analysis. Finally, we suggest changes that would improve the quality of data synthesis in ecology and evolutionary biology, and predict future directions for this emerging enterprise.

II. INTRODUCTION

If one browses through ecology journals published before the 1960s, it becomes apparent that statistical tests were not commonly employed in ecological research prior to that time. The integration of formal statistical methods into scientific practice subtly but profoundly changed the perspective of ecologists. While anyone trained in modern ecology is now familiar with at least basic statistical methods for quantifying and comparing responses within a study, many are not yet fully aware of dramatic new developments for the quantitative synthesis of research results across independent studies. Modern statistical techniques for quantitative research synthesis, collectively known as *meta-analysis*, were

first introduced to the fields of ecology and evolution in the early 1990s (Table 1 and Figure 1; Arnqvist and Wooster, 1995a). Glass introduced the term meta-analysis, defining it as 'the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings' (Glass, 1976, p. 3), contrasting it with the primary analysis of individual studies. A more recent definition that emphasizes contemporary approaches is that meta-analysis is a set of 'statistical methods designed to draw rigorous inferences from multiple studies' (H. Caswell, personal communication). Meta-analysis is typically used to synthesize the results of published studies, but it may also serve to combine published or unpublished results by an individual investigator or research group, or by a group of investigators who have agreed, either in advance or after the work is completed, to combine their results. We contrast
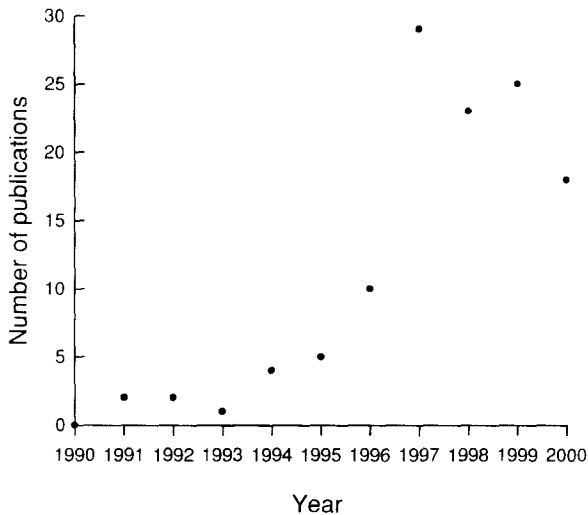
Fig. 1. Number of papers published per year on meta-analysis in the fields of ecology and evolution. The list (references are in Table 1) is inclusive and non-selective, and contains, in addition to strictly defined meta-analyses, papers on meta-analysis methodology, comments, reviews and critiques of meta-analyses in these fields. Several papers claim to be meta-analyses but do not meet the definitions laid out in the text of this review. The purpose of being more inclusive rather than more selective was to gather all of the literature on the topic to date for other researchers interested in various aspects of the application of meta-analytic techniques in these research fields: further sorting and selection is therefore available to the reader. We attempted to gather references to all recent quantitative syntheses and discussions of meta-analysis applications by searching the Web of Science (ISI) using keywords, references to key papers, and papers cited in other references. All papers were then reviewed for their appropriateness for inclusion in this review.

Table 1
Papers on meta-analysis in ecology and evolutionary biology included in Figure 1

Year	Authors	Year	Authors	Year	Authors
1991	Jarvinen Marchant and McGrew	1997 (cont.)	Hugueny and Guegan Kraak Leamy	1999	Arft *et al.* Byers and Waller Downing *et al.* Englund and
1992	Gurevitch *et al.* VanderWerf		Liermann and Hilborn Markow and Clarke		Evander Englund *et al.*
1993	Gurevitch and Hedges		McCarthy McGrew and		Goldberg *et al.* Gurevitch and Hedges Hedges *et al.*
1994	FernandezDuque and Valeggia Poulin Tonhasca and Byrne Wooster		Marchant Miller and Fair Møller and Thornhill (a, b) Myers *et al.* Osenberg *et al.*		Medlyn *et al.* Møller Møller and Alatalo Møller and Shykoff Osenberg *et al.* (a, b)
1995	Arnqvist and Wooster (a,b) Peterman Turchin Vasquez *et al.*		Palmer and Strobeck Pomiankowski Schalk and Forbes Swaddle Thrush *et al.* Whitlock and Fowler		Palmer Petersen *et al.* Peterson *et al.* Planque and Fredou Riessen Simmons *et al.*
1996	Arnqvist *et al.* Brett and Goldman Britten Csada *et al.* Curtis Dahl and Greenberg Ellison Frankham Leung and Forbes Poulin	1998	Bender *et al.* Blackenhorn *et al.* Cebrian *et al.* Curtis and Wang Fiske *et al.* Folt *et al.* Hartley and Hunter Herrera *et al.* Hilborn and Liermann	2000	Thornhill *et al.* Van Dongen *et al.* Vollestad *et al.* Wand *et al.* Xiong and Nilsson Arnqvist and Nilsson Brook *et al.*
1997	Abouheif and Fairbairn Adams *et al.* Bauchau Brett and Goldman Brett and Mueller- Navarra Cote and Sutherland Dodds Dolman and Sutherland FernandezDuque Hamilton and Poulin Hechtel and Juliano Houle		Koricheva *et al.* (a,b) Lardicci and Rossi McCann *et al.* Møller Møller and Ninni Møller and Thornhill Murray Myers Myers and Mertz Piegorsch *et al.* Proulx and Mazumder Van Zandt and Mopper Vernier and Fahrig Westoby		Collie *et al.* Connor *et al.* Gliwicz and Glowacka Gough *et al.* Gurevitch *et al.* Hollister and Webber Jones Lempa *et al.* Leung *et al.* Mosquera *et al.* Pither and Taylor Poulin Rosenberg *et al.* Schmitz *et al.* Sokolovska *et al.* Windig and Nylin

contemporary meta-analysis with other approaches to reaching conclusions from a body of data obtained from a set of independent studies in Section IV.

Meta-analysis can address many of the inferential issues familiar from classical statistics, albeit at a differ.ent level (i.e. across studies), including parameter estimation (point estimates and confidence intervals) and hypothesis testing, using parametric or non-parametric models. In addition, it can provide additional information not offered by classical statistical methods, particularly about the consistency of the results across studies. The synthesis of research results is a universal and venerable occupation in science (see, e.g. Cooper and Hedges, 1994b, for a lucid overview of this topic); indeed, scientific papers almost always begin with putting their results into the context of what else is known about the topic. Meta-analysis provides tools to make this process quantitative and more rigorous, just as classical statistics allowed rigorous quantitative inferences to be made about the results of single experiments.

This review is aimed at a range of ecologists, from those who may have heard the term meta-analysis, but are not sure what exactly it entails, to those who have read a particular meta-analysis paper and wish to put it into a more general context, to those who may be undertaking a meta-analysis themselves. Our goals in writing this chapter are to introduce what meta-analysis is and is not; to review briefly its precedents; to identify all publications concerning meta-analysis in ecology and evolution to date and to highlight the recent history of this topic; to consider some of the controversies concerning its application in this field and consider some of the difficulties that arise in the synthesis of multiple studies; and, finally, to illustrate its application in ecology with several recent meta-analyses in the area of global change biology.

Meta-analysis can be viewed as a tool that enables investigators to see a larger picture, one that is not apparent when looking at some, or even all, of the component studies individually. In this sense, meta-analysis resembles a synthetic, composite photograph in which large numbers of small, individual photos of many different subjects are arranged to create a single, unique larger image that comes into focus only when seen from a greater distance. In this analogy, each of the small individual photos represent the individual studies or data records, and the composite picture is the result of meta-analysis. We suggest that the adoption of these techniques in the discipline of ecology will transform the way research results are understood, much as meta-analysis has in other disciplines, and with an impact potentially as great as that of the introduction of statistical analysis in ecology in the middle of the twentieth century.

Meta-analysis raises philosophical issues regarding the ability to reach a general understanding based upon a body of data from independent

studies. The issue of generalizing research results is a particularly ambiguous and troubling one in the field of ecology, where we seek to understand natural systems, and yet work with such a great diversity of organisms, systems and levels of organization. On the one hand, most ecologists, like other scientists, are trained in graduate school to be cautious about over-extrapolating their results. As a consequence, they may confuse scientific rigor with conservatism, hesitating to extend their conclusions to other organisms or conditions. On the other hand, we all begin learning formal ecology from textbooks, which often use the results of particular experiments as examples to illustrate or prove the existence of important phenomena. Perhaps as a result, ecological thinking can demonstrate a curious kind of naivete in accepting individual experimental results as if they constituted the entire truth about nature – the habit of accepting 'textbook examples' (or alternatively, the most recently published study) as if a single experimental result could conclusively prove the reality of a particular ecological phenomenon, or tell us all we might ever want to know about it. Somewhere between the opposite extremes of hesitating to extend the implications of particular results, and over-generalizing from exemplary experiments, ecologists and evolutionary biologists need to consider more deeply the issue of how we use experimental evidence to make generalizations about nature.

Ecological studies, while highly diverse in methodology, subjects and systems, often address questions of broader interest that may be relevant to other species, systems and settings. Skeptics of ecological meta-analysis question the validity of obtaining any quantitative summary of the results of a diverse body of experimental data, owing to the inevitable inadequacies in the primary data as well as to perceived shortcomings in meta-analytic techniques. But there is nothing out of the ordinary in wishing to summarize the findings of a group of studies, as is routinely done in review series such as this one. What is different in meta-analysis is that a formal methodology has been developed to quantify syntheses and test hypotheses about the overall results.

How can the results of different experiments be synthesized quantitatively in a rigorous way? First, we must define what we are attempting to accomplish in a quantitative synthesis. In summarizing the available evidence about an effect (e.g. the effect of predation in field studies) in which we are interested, some of the important questions one might wish to answer are: How large is the overall effect (does predation have a large effect, a small effect, or no detectable effect on prey density across studies)? Can we be confident that this effect is greater or less than zero (is the estimated value of the effect significantly different from zero, when sampling variance is taken into account)? Are the results consistent across studies or, if not, are there systematic differences among categories of

studies in the magnitude of the effect (e.g. among trophic levels, or for studies of long versus short duration)? Meta-analysis statistics provide straightforward means by which such questions can be answered. We next turn briefly to the basic statistical methods for meta-analysis, and examine why contemporary meta-analytic methods are superior to the more familiar alternatives for synthesizing the results of multiple studies.

III. STATISTICAL APPROACHES IN ECOLOGICAL META-ANALYSIS

A. Scaling Responses across Studies Using Metrics of Effect Size

Conducting a meta-analysis requires that the results of separate studies are put onto a common scale, so that they can be compared and averaged. Glass' insight into one way in which it could be done signalled the beginning of modern meta-analysis (Cooper and Hedges, 1994b). One of the most common approaches to combining studies in contemporary meta-analysis is to standardize the outcomes using some metric of *effect size*. Various effect size metrics have been proposed, and some that are commonly used in meta-analysis are listed in Table 2. We discuss the issue of the choice of effect size metrics below.

Two commonly used metrics in ecology, the standardized mean difference, Hedges' d, and the log response ratio, lr, are typically used where the outcome expresses the magnitude of the response to an experimental treatment by comparing an experimental to a control group. For example, it might be meaningful to say that the mean mass of organisms in the experimental group was one standard deviation greater than that in the control group (i.e. $d = 1.0$), or that it was 1.25 times the mass in the control group (i.e. $lr = \ln 1.25$; Figure 2). In some research areas the outcomes are reported as correlation coefficients, and these can be used as a metric of effect size in a meta-analysis (because Z has superior statistical properties, correlation coefficients are transformed using Fisher's Z-transform before carrying out the analysis, Table 2).

In experiments in which the results are categorical, a common metric taken from medical meta-analysis is the odds ratio (or other members of the family of measures to which it belongs; see Rosenberg *et al.*, 2000). Imagine that responses are presented in the form of 2×2 contingency tables. For example, we might be interested in whether or not seeds were removed by dispersers (i.e. the outcome is expressed as numbers removed/not removed), when seeds were experimentally colored red or left uncolored (colored/uncolored). The numbers in each cell of the contingency table are expressed as in Table 3. Following the notation in

Table 2
Summary of some common effect metrics and their sampling variances

Metric	Symbol	Equation		Sampling variance
Hedges' d	d	$d = \dfrac{(\bar{X}^E - \bar{X}^C)}{S}J$	$v_d = \dfrac{N^C + N^E}{N^C N^E} + \dfrac{d^2}{2(N^C + N^E)}$	
Response ratio	lr	$lr = \ln\left(\dfrac{\bar{X}^E}{\bar{X}^C}\right)$	$v_{lr} = \dfrac{(s^E)^2}{N^E(\bar{X}^E)^2} + \dfrac{(s^C)^2}{N^C(\bar{X}^C)^2}$	
Correlation coefficient – Fisher's z-transform of r	z	$z = \dfrac{1}{2}\ln\left(\dfrac{1+r}{1-r}\right)$	$v_z = \dfrac{1}{N-3}$	
Odds ratio relative odds	$\ln OR$	$OR = \dfrac{P_t(1 - P_c)}{P_t(1 - P_t)}$	$v_{\ln OR} = \dfrac{1}{T} + \dfrac{1}{C} + \dfrac{1}{t} + \dfrac{1}{c}$	

Symbols are as follows: $\bar{X}^E$ and $\bar{X}^C$ are the means of the experimental and control groups, respectively, S is their pooled standard deviation, and N^E and N^C are their respective sample sizes; the term J corrects for small-sample bias (Hedges and Olkin, 1985). Pearson's correlation coefficient, r, is z-transformed, with its associated sample size, N. The odds ratio is explained in the text.
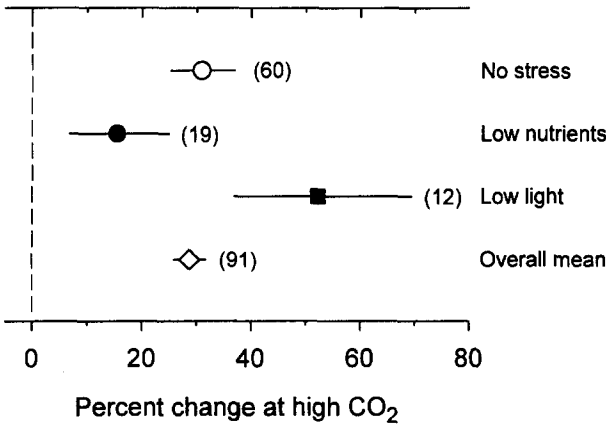
Fig. 2. Results of a meta-analysis of the effects of environmental stress on the biomass responses of trees to elevated CO_2. Experimental treatments were: no stress, low nutrients, or low light levels. Means $\pm 95\%$ confidence intervals (CIs), with number of studies in parentheses. (Data from Curtis and Wang, 1998.)

Table 3, the rate of response for the treatment group, P_t (seeds colored red), is

$$P_t = \frac{T}{n_t}$$

where T is the number of seeds exhibiting a response ('removed') in the treatment group, and n_t is the total number (responders and non-responders) in the treatment group. The rate of response for the control group, P_c (uncolored seeds), is

$$P_c = \frac{C}{n_c}$$

where C is the number of seeds exhibiting a response ('removed') in the control group, and n_c is the total number (responders and nonresponders) in the control group. The odds ratio and its sampling variance is given in Table 2. Similarly, responses in other ecological studies could fit a wide variety of possible categorical outcomes (dead/alive, mated/unmated, metamorphosed/remained larval, etc.), and although this metric has been used more commonly in medical research, it is appropriate in many cases of ecological research as well.

The choice of effect size metric can potentially have a substantial impact on the results of a meta-analysis. Osenberg and colleagues (Osenberg et al., 1997, 1999b; Osenberg and St Mary, 1998) have criticized the use of standard effect size metrics in ecology and evolution. Instead, they suggest that each time a meta-analysis is conducted, the authors should model the biological process being studied as well as the spatial/temporal scales of the experiments, and, based upon the model used, construct appropriate effect size metrics. They criticize the uncritical adoption of standard effect size metrics in meta-analysis, particularly the use of d. A major criticism of d that they raise is based upon the possibility that d may confound the differences between treatment means with what are effectively artifactual

Table 3
2 × 2 contingency table (see text)

	Treatment	Control	Total
Removed	T	C	$T + C$
Not removed	t	c	$t + c$
Total	$n_t = T + t$	$n_c = C + c$	$T + C + t + c$

differences in variance among studies (e.g. due to experimental design or spatial heterogeneity). This echoes the argument made by Hurlbert (1994) criticizing the use of d in ecological meta-analysis. In other words, the number of standard deviations by which the experimental and control group means differ is, in such cases, not an accurate reflection of the true magnitude of the effect being studied. It would also be true, in such cases (but not discussed by Osenberg and colleagues) that in the primary analysis of each of such studies, standard statistical analyses such as ANOVA (analysis of variance), regression or t-tests would generally be inappropriate as well. They also feel that d would not effectively capture the biologically most meaningful measure in combining the results of many ecological studies.

We could not disagree more strongly with the arguments of Osenberg and his colleagues that different effect size metrics should be used each time a meta-analysis is carried out, although we enthusiastically second their recommendation that meta-analysts think carefully about their analysis (including the choice of effect size metric and the interpretation of the results). If a comparison between the experimental and control groups is not effectively expressed in terms of standard deviation units because there are systematic, artifactual differences in the magnitude of the variances among studies, a ratio-based metric such as *lr* (Hedges *et al.*, 1999) should overcome that problem.

There are two major objections to creating new metrics for each meta-analysis: first, lack of standardized measures makes evaluation of the results of the meta-analysis very difficult, and second, for most ecologists, determining the statistical properties of novel statistics is essentially out of reach. The use of standard metrics means that anyone reading a meta-analysis can understand the results, and can compare results among different meta-analyses. This is analogous to why we have uniformly adopted the SI system for biological measurements, rather than creating the biologically most meaningful metric based upon constructing a conceptual or mathematical model each time we measure an organism or field plot. One can run into serious and largely invisible statistical potholes in using novel metrics with unknown sampling distributions and unknown properties, and the use of conventional meta-analysis methods for calculating means, confidence intervals, conducting homogeneity tests, and other analyses of interest on such metrics is likely to result in misinformation. A more minor limitation to the suggestion of Osenberg and co-workers is that for most meta-analyses in ecology there will not be a single way to model all studies one is interested in combining, and it is unlikely that they will all be on the same spatial or temporal scale except for meta-analyses of unusually limited scope. There is an alternative to creating new metrics for each meta-analysis on the one hand, and on the other having a

single metric which is sometimes highly unsatisfactory for particular ecological data. That is to have a body of well-understood metrics (perhaps, say, a half-dozen) to choose from that represent a range of biologically meaningful types of comparisons between groups. This may represent a satisfactory compromise between the two positions for most ecological meta-analyses.

B. Combining Results across Studies

Once one has calculated an appropriate effect size for each study, these measures of response can be combined in several ways. There are many advantages to using weighted analyses (see Gurevitch and Hedges, 1999; Hedges and Olkin, 2001), in which effect sizes are weighted by the inverse of their sampling variances. The weights are typically chosen to provide unbiased estimates with minimum variance. One might first determine the weighted grand mean effect size, $\bar{E}$, where

$$\bar{E} = \frac{\sum_{i=1}^{n} w_i E_i}{\sum_{i=1}^{n} w_i}$$

n is the number of studies, and E_i is the effect size for the ith study. The weight for the ith study is the reciprocal of its sampling variance, $w_i = 1/v_i$ (see Table 2). The variance of $\bar{E}$ is:

$$s_{\bar{E}}^2 = \frac{1}{\sum_{i-1}^{n} w_i}$$

Using $s_{\bar{E}}^2$, the confidence interval around $\bar{E}$ is:

$$CI = \bar{E} \pm t_{\alpha/2[n-1]}(s_{\bar{E}})$$

where t is the two-tailed critical value found from the Student's t-distribution at the critical level, α, at $n-1$ degrees of freedom. The grand mean effect size thus presents one with a means for assessing the magnitude of the effect across all studies, and its confidence interval offers a way to evaluate whether the effect overlaps zero or is significantly greater or less than zero.

C. Explaining Differences in Response among Studies

Beyond the bare bones of obtaining a mean effect across studies and the confidence interval around that mean, one might wish to ask, are the results consistent across studies, and if not, can we explain differences in the effect beyond those due to random sampling error? The appropriate analysis will depend, of course, upon the nature of the questions one is posing, and on the nature of the data. Categorical (ANOVA-analog) and continuous (regression-analog) approaches have both been developed for meta-analytic data, although the first has been used more commonly in ecological and evolutionary applications. Both depend upon weighting effect sizes by the inverse of the sampling variances. Categorical approaches rely upon homogeneity tests. The body of results is tested to determine whether they are consistent with one another across studies, after accounting for sampling error, or whether there is genuine heterogeneity in the magnitude of the effect among studies. If studies are heterogeneous, it is often of interest to test whether there are consistent differences among categories of studies that are hypothesized to respond in different ways (carnivores versus herbivores, long-term versus short-term studies, etc.). Continuous models are useful where the effect sizes are hypothesized to depend upon an independent variable (e.g. productivity, Goldberg *et al.*, 1999; Figure 3).
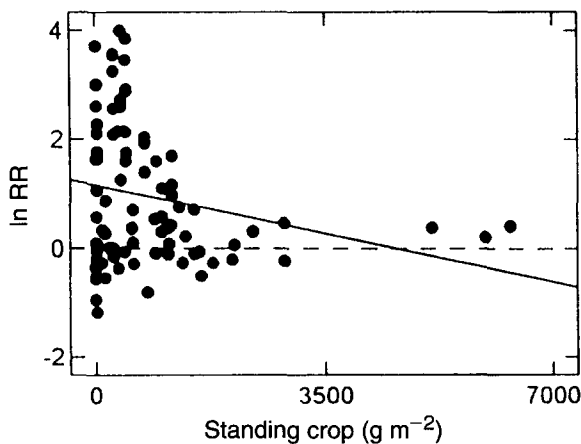
Fig. 3. Effect of competition for plants (expressed as log response ratio, RR, of plant biomass with reduced competition compared to that with competitors present) as a function of standing crop in the community. Reproduced with permission from Goldberg *et al.* (1999), Figure 1D.

One of the most important distinctions in the type of analysis for both categorical and continuous data is between fixed and random effects analyses. In simple fixed effects models it is assumed that all studies share a common 'true' effect size, and differ only by sampling error. In contrast, random effects models assume that there is random variation among studies in the 'true' effect. Mixed models combine random variation among studies within a category with fixed differences among categories (Gurevitch and Hedges, 1993, 2001). Fixed versus random effects models differ in these and other assumptions, and their results have somewhat different interpretations (e.g. Gurevitch and Hedges, 1993, 1999, 2001; Raudenbush, 1994). The methods for data analysis based upon fixed, random and mixed models for categorical and continuous data can be found in textbooks (e.g. Hedges and Olkin, 2001) and are available in some software packages (e.g. Rosenberg et al., 2000).

Other statistical approaches in meta-analysis include the use of resampling tests (randomization and bootstrapping) for both continuous and categorical analyses, where parametric tests may not be appropriate (Adams et al., 1997; Rosenberg et al., 2000). A subtle difference between standard parametric approaches and randomization tests concerns the implications regarding the universe to which one is generalizing. Parametric tests are based upon the perspective that the studies in the meta-analysis represent a population of experimental results that *could* have been done, and therefore the meta-analysis results apply to some broader population of studies beyond the actual ones included in the synthesis. Randomization tests, in contrast, are sometimes seen as viewing the data set in hand as the entire universe, with the conclusions legitimately applying only to those studies. The latter perspective may be easier to justify, at least in meta-analyses in the field of ecology. Another recent development is a technique for factorial meta-analysis, where the studies being synthesized have all manipulated the same two treatment factors at each of two levels (e.g. competition and predation; Gurevitch et al., 2000). This technique allows one to evaluate not only the main factors, but also their interaction across studies. Bayesian methods have also been applied to meta-analysis in ecology and evolutionary biology (Ellison, 1996; Liermann and Hilborn, 1997; Hilborn and Liermann, 1998; Myers and Mertz, 1998; Piegorsch et al., 1998).

IV. LIMITATIONS TO TRADITIONAL METHODS FOR DATA SYNTHESIS

The most traditional method for synthesizing the results of a body of studies is the narrative review, in which an expert in the field provides

an overview, evaluation and conclusions based upon the literature on a particular question or topic. While meta-analysis offers an alternative to the traditional narrative review, it will hardly make *Advances in Ecological Research* and other journals that publish narrative reviews obsolete. Narrative reviews serve many valuable functions, and continue to be published in those disciplines in which meta-analysis holds a firmly established role. Some of the things narrative reviews can offer are the authors' perspective on the status of a discipline, evaluation of current techniques and the state of the art, and provocative questions and directions for future research. But narrative reviews also suffer from some inherent flaws. They are almost unavoidably subjective, both in their choice of studies from those available and in their conclusions (e.g. Light and Pillemer, 1984; Cooper and Dorr, 1995; Peipert and Bracken, 1997; Davidoff, 1999; Swales, 1999). Even when reviewers try their best to avoid subjectivity and bias, the selection of studies and the evaluation criteria often are eccentric or haphazard. The impressions of the reviewers are too often based largely upon the Abstract and Discussion sections of papers rather than on detailed examination of the actual data. Most importantly, narrative reviews cannot provide the information that a meta-analysis can. In particular, only a quantitative summary of the results can answer questions such as: What is the magnitude of the response overall? Is it consistent across studies or not? What explanatory variables might account for differences in the magnitude of the response among studies?

A second conventional (if less venerable) approach to research synthesis in ecology can be called 'vote-counting'. In attempting to transcend the limitations of strictly narrative reviews, over the past two decades ecologists unfamiliar with the literature on meta-analysis in other disciplines have devised at least two basic quantitative approaches to synthesize research results. The most common of these approaches is simply to count the number of statistically significant and non-significant outcomes and then weigh these counts against each other (sometimes with the use of statistical tests) to determine the magnitude and importance of the effect of interest. This might entail, for example, attempting to evaluate the existence and magnitude of the effect of competition in field experiments, with the assumption that the importance of the effect will be reflected by having it detected by a large proportion of the studies testing for it. This approach is known as vote-counting, as if each study casts a vote for or against the effect, and it has been used in many other fields as well as in ecology. However, vote-counting has such serious flaws and limitations that it is essentially never advisable to employ it. In other disciplines, these flaws have become well known and are generally widely understood, and vote-counts are no longer published. In addition, there are more subtle limitations to vote-counting that are at least in part

ameliorated by using meta-analysis, as explained below. Unfortunately, vote-count reviews continue to be published in ecology (see, e.g. Gurevitch and Hedges, 1999), although the problems with this approach have been pointed out in this discipline for the better part of a decade (e.g. Gurevitch and Hedges, 1993).

The central problem with vote-counting concerns statistical power. The statistical significance of the outcome of an experiment depends both on the magnitude of the observed response, on the variability in the data, and on the sample size, which together determine the likelihood that the experiment will succeed in detecting a real response. This results in a negative bias, because the ability to detect real effects (i.e. to reject the null hypothesis) is limited when sample sizes are small and the magnitude of the effects are modest. (By negative bias, we mean that there is a bias against detecting true effects, because we fail to reject the null hypothesis too often when in fact it should be rejected.) Narrative reviews are subject to the same problem, because the reviewer often relies upon the statistical significance of the results in evaluating the outcome of the studies being reviewed. Because ecological studies typically have small sample sizes and modest effects, they are particularly vulnerable to this negative bias. It is also common to decide whether two studies agree or disagree based upon the statistical significance of their outcomes, but when statistical power is low, even studies with identical underlying effects have a fairly large chance of disagreeing by chance alone (e.g. Gurevitch and Hedges, 1999). In statistical terms, vote-counting is not a very good procedure: it has low power itself to detect results, the conclusions drawn are highly biased, and, counter-intuitively, its statistical power actually declines as the number of studies reviewed increases (the reliability of the test actually becomes worse as one gains more information; Hedges and Olkin, 1980, 1985). Readers should note, however, that there is a statistically robust variant of vote-counting that may be of use when the available data are extremely limited (Hedges and Olkin, 1985). Vote-counting, like narrative reviews, cannot provide reliable information on the magnitude or consistency (homogeneity) of the results of the studies being reviewed. Standard techniques for meta-analysis are not subject to these problems and limitations.

A more subtle limitation with vote-counting that is more likely to be resolved in carrying out a meta-analysis is the lack of specificity in what one is trying to answer in synthesizing results. The 'effect' being evaluated is often only vaguely specified, in contrast to the well-defined 'effect size' used in meta-analysis. In attempting to determine the importance of an effect, vote-count reviews often do not distinguish between statistical significance, biological significance, and biological importance, and confuse the frequency with which effects are detected with their magnitude and

importance. In evaluating the statistical significance of each study, since each study used different statistical models and different treatments, it is not so clear what the statistical significance being evaluated is, nor what the effect is whose magnitude is being considered. In conducting a meta-analysis, each of these and many other issues must be carefully weighed, and a decision must be reached and justified. These decisions should be specified in the published meta-analysis.

Another conventional, but flawed, approach to quantitative research synthesis is to calculate standardized effect sizes, but to use more familiar statistical techniques like ANOVA and ordinary least-squares regression to carry out quantitative syntheses of the effect size estimates across studies. This is not advised, and we briefly review the reasons below. Presumably, these reviewers acknowledge the utility of using standard effect size measures (as introduced above) for integrating results across studies, and may even be convinced of the inappropriateness of vote-counting, but are unfamiliar with the statistical techniques used in meta-analysis. The issues involved are spelled out in more detail by Gurevitch and Hedges (1999) and Hedges and Olkin (2001).

Population estimates based on large sample sizes are more precise than those based upon small sample sizes. Because studies in a meta-analysis data set may differ by orders of magnitude in sample size, sampling variances also commonly vary enormously in such data sets. For this reason, concern about violating the usual statistical assumption of homogeneity of variances is often much more serious in meta-analysis than in primary statistical analyses. Data transformation cannot eliminate this problem, and even very high levels of heteroscedasticity may be cryptic, escaping the notice of the analyst who only looks at effect sizes but not at their sampling variances. ANOVA and ordinary least-squares regressions may perform very poorly (i.e. give inaccurate results) in such circumstances. Standard meta-analytic techniques avoid this problem because the analyses explicitly take sampling variance into account.

To illustrate this problem, consider that in a conventional primary analysis, data collected on individuals (organisms, plots, etc.) can be represented by points on a graph. The graph could be a scattergram, for example, where responses of individuals are hypothesized to be dependent upon some other variable. Alternatively, the points might merely illustrate the degree of scatter about the mean. These individual points are known with a fair degree of certainty, because they have been measured directly (there may be a small amount of measurement error associated with them, but we can assume that that will be minor relative to other sources of variation). In a meta-analysis, the data are no longer measurements on individuals known with some certainty, but rather are estimates of population parameters (such as the effect sizes in each study). As in a

primary analysis, the researcher might be interested in relating explanatory variables to the outcomes (which here are effect sizes). One might picture these meta-data not as points on a graph, but rather as clouds on a graph. Each cloud (the effect sizes in each study) will be different in size from the others, because these are population estimates, which are known with greater or lesser certainty in each study. The magnitude of each 'cloud' is inversely proportional to its sampling variance (Table 2); studies with large sample sizes approach points (i.e. their effect sizes are known fairly precisely), while those with small sample sizes are large clouds with a lot of uncertainty as to where the actual effect size is located in the cloud. ANOVA, regression and other classical statistical methods are not robust to violations of the assumption that this sampling variance is equal among points (i.e. studies). Again, meta-analysis avoids this problem by explicitly taking sampling variance into account.

V. A BRIEF HISTORY OF META-ANALYSIS IN ECOLOGY AND EVOLUTIONARY BIOLOGY

A. Origins of Contemporary Meta-analysis

While modern methods for meta-analysis have been developed largely in the last quarter century, these techniques have old roots, traceable to methods that include the combined tests of statistical significance developed by such luminaries as Cochran (1937), Fisher (1932), Pearson (1904, 1933) and Tippett (1931). The modern methods, however, are both more powerful and much more useful than those developed earlier.

The development of formal techniques and approaches for modern quantitative data synthesis began in the 1970s in the social sciences, particularly in clinical, educational, social and industrial psychology (see Cooper and Hedges, 1994b; Hunt, 1997). In the following decade, as researchers became aware of these tools and the statistical techniques available were further developed and refined, there was an explosion of interest in meta-analysis in medical research (e.g. Sacks *et al.*, 1987; Chalmers *et al.*, 1989; Mann, 1990, 1994). Cooper and Hedges (1994a, p. 7) state that:

> Research synthesis in the 1960s [in the social sciences] was at best an art, at worst a form of yellow journalism. Today, the summarization and integration of studies is viewed as a research process in its own right; it is held to scientific standards and applies the techniques for data gathering and analysis developed for its unique purpose.

These 'near-revolutionary developments' (Cooper and Hedges, 1994b) in research integration experienced in the behavioral and medical sciences

over the past 20 years have only recently begun to have an impact on the field of ecology. As the discipline earlier made the transition from description and classification to modern quantitative approaches, research synthesis in ecology is poised to make the transition from descriptive to formal quantitative techniques. Ecologists in earlier periods generally possessed an intimate knowledge of the biology and natural history of the organisms they studied and about the environments in which they were found. The richness of detail that was once common in describing individual organisms, or the alternative, the imperative to classify and categorize organisms and systems, eventually gave way to an emphasis on representing (statistical) populations by means and variances, and comparing them with other populations using statistical hypothesis testing. The introduction of statistical methods into ecological and evolutionary research was initially strongly resisted in some quarters (see Preface to Simpson *et al.*, 1960), but ultimately changed the very nature of the science. This change coincided with many others in the discipline, and was basic to completing the transition of ecology from natural history to a modern scientific discipline. The fundamental shift in outlook that this entailed may be echoed, in some ways, by that resulting from the gradual acceptance of modern methods for the statistical synthesis of research results in ecology.

B. Publications in Ecological Meta-analysis

We attempted to collect all papers concerned with formal meta-analysis in ecology and evolution. We conducted an exhaustive search, relying primarily on the ISI Web of Science on-line scientific publication database. We searched for papers by keywords and by citations of known key references in the field. We then examined the articles found to determine if they were appropriate. It was our intention to be highly inclusive; papers were not screened for quality, proper use of statistical methods, or other selective criteria. Both actual meta-analyses as well as papers concerned with meta-analysis methodology, application and controversies were included in our comprehensive list. The purpose of using very broad selection criteria was to collect all potentially useful papers for future use by readers interested in this topic, who might choose to screen the papers we found using different criteria of their own.

   The first meta-analysis in ecology was published in 1991, and the number of publications concerned with the use of this approach has continued to increase since then (Figure 1). We identified 119 publications concerned with meta-analysis in ecology and evolution through 2000. While it is likely that the slight decline in number of publications in 2000 is due to random variation, the trajectory in the future obviously remains to be seen. Although it is clear that the volume of publications in meta-analysis in

ecology and evolutionary biology will never equal the hundreds of meta-analyses published in medicine each year, there are a number of reasons to suggest that meta-analysis is beginning to become firmly established in this field, as we will illustrate. From its introduction, meta-analysis has been applied – and debated – in a wide range of subdisciplines in the field, rather than being restricted to any particular area (as illustrated by the subjects touched upon in the titles of the papers cited in Table 1). Some of these efforts have been inspired by a few influential papers, but the diversity of applications also suggests the independent discovery of these techniques by various authors at the same time that experimental data in a variety of subdisciplines within ecology surpassed a critical mass during the 1990s.

Meta-analysis has been used for the analysis of the results of different experiments by an individual or group (e.g. Hechtel and Juliano, 1997), as well as for sweeping syntheses of the literature on a topic (e.g. Myers and Mertz (1998) included more than 500 fish populations, while Curtis and Wang (1998) evaluated over 500 reports of effects of elevated CO_2 on woody plants). Organisms serving as the focus of ecological or evolutionary meta-analyses have included plants (from trees, e.g. Curtis, 1996; to seagrass, Cebrian et al., 1998), apes (e.g. Marchant and McGrew, 1991), fish (e.g. Myers et al., 1997; Folt et al., 1998) insects (e.g. Tonhasca and Byrne, 1994; Koricheva et al., 1998a) and birds (e.g. Jarvinen, 1991; Cote and Sutherland, 1997), as well as broad arrays of organisms (e.g. aquatic food webs, Brett and Mueller-Navarra, 1997; competing populations, Gurevitch et al., 1992; and lekking species, Fiske et al., 1998). Meta-analysis has been used to address problems ranging from assessing issues of applied importance in conservation ecology (e.g. Bender et al., 1998; Hilborn and Liermann, 1998) to evaluating the strength of evidence for phenomena predicted by evolutionary theory (e.g. Britten, 1996; Møller and Thornhill, 1998). Other papers listed in Table 1 include those that have made suggestions for applications of meta-analysis in ecology, promoted or questioned its usefulness and validity, and developed new methodology for its use.

C. Reviews, Symposia and Other Activities

The first general review of meta-analysis in ecology was published in 1995 (Arnqvist and Wooster, 1995a); this is the second. In addition to publications, various other activities concerning meta-analysis in the field of ecology have occurred in recent years. In 1996, a symposium was held at the Ecological Society of America's annual meeting entitled 'Meta-analysis in Ecology' that sought to introduce meta-analysis more broadly to the ecological community and demonstrate its range of applicability (Brett, 1997). Symposium presentations ranged from overviews of available

techniques for meta-analysis to case studies of applications of meta-analysis to ecological problems. Participants identified a range of issues of importance with respect to the wider use of meta-analysis in ecology. A software package (MetaWin 1.0, Rosenberg *et al.*, 1997), specifically targeted at ecological meta-analysis, was introduced at this symposium.

Also in 1996, a working group was convened at the National Center for Ecological Analysis and Synthesis (Santa Barbara, USA) to help evaluate and guide the application of meta-analysis to ecological questions (Osenberg *et al.*, 1999a). Through a series of workshops, participants examined the linkage between effect size metrics and ecological models, evaluated the statistical properties of these metrics, and conducted meta-analyses of published data to address a broad range of ecological problems, including plant community ecology (Goldberg *et al.*, 1999), predator–prey interactions in stream ecosystems (Englund *et al.*, 1999), and marine nutrient cycling (Downing *et al.*, 1999). A common thread in this work was the recognition of the unique features of ecological data as distinct from, say, psychological or medical data. This critical point was further developed by Osenberg *et al.* (1999b) who examined the nature of hypothesis testing in ecological meta-analysis, and by discussions of statistical methods tailored to ecological applications by Gurevitch and Hedges (1999) and Hedges *et al.* (1999).

It is fair to say that currently the awareness and understanding of meta-analysis among ecologists and evolutionary biologists is still highly uneven. Methods used for recent quantitative syntheses in ecology and evolutionary biology have also spanned a gamut of approaches. Some authors have developed and applied sophisticated extensions of meta-analysis statistical techniques, taking into account the peculiarities of the data sets with which they worked, while others continue to use vote-counts, or even more curiously, combine meta-analysis methods with vote-counting. Ecological journals continue to publish inappropriate use of ANOVA, goodness-of-fit tests, and other methods designed for primary data analysis where meta-analysis methodology should be applied, as well as vote-counts that call themselves meta-analyses and others written by authors who are apparently unaware of meta-analysis (e.g. see Gurevitch and Hedges, 1999). Sometimes such papers have even been published back-to-back with sophisticated applications of meta-analysis.

D. Examples of Several Recent Controversies

One of the more controversial applications of meta-analysis in this field has been to evaluate evidence for a hotly debated area in evolutionary theory, fluctuating asymmetry and developmental instability (Møller and Thornhill 1997a,b, 1998; Thornhill *et al.*, 1999). Fluctuating asymmetry

refers to deviations from symmetry, typically in animal morphology. Fluctuating asymmetry is thought to represent one form of developmental instability, which is a reflection of the ability (or inability) of individuals to undergo stable development under given environmental conditions. Controversy swirls around many aspects of this topic, including whether it is a genetically based trait (i.e. is heritable), and whether it is under sexual selection (i.e. more asymmetric individuals are at a disadvantage in attracting mates). Møller and Thornhill (1997a,b) integrated heritabilities of individual fluctuating asymmetry from 34 studies of 17 species and found that across all studies, heritability was significantly greater than zero. They concluded that this provided evidence for the existence of an additive genetic component to developmental stability. Seven commentaries on and critiques of Møller and Thornhill's meta-analysis were published (Houle, Leamy, Markow and Clarke, Palmer and Strobeck, Pomiankowski, Swaddle, Whitlock and Fowler, as well as a reply by Møller and Thornhill, in the *Journal of Evolutionary Biology* (Møller and Thornhill, 1997b)). Many of these authors took the opportunity to discuss the general merits of meta-analysis, in addition to commenting on the validity of the specific conclusions drawn. One point of concern among a number of the reviewers was the combination of effect sizes drawn from studies of either differing quality or, more importantly, from different species and on different morphological traits. These concerns reflect the 'garbage-in, garbage-out' and 'apples versus oranges' problems, respectively, that we consider in more detail below (see sections VI.D and VI.F). Others questioned the objectivity of the meta-analysis. The spirited debate clearly demonstrated the emergence, and controversial nature, of meta-analysis as a new tool for testing evolutionary theory. Further papers by Møller, Thornhill and colleagues, and by their critics, have continued the debate on both fluctuating asymmetry and on the use of meta-analysis in evolutionary research (Palmer, 1999, Thornhill *et al.*, 1999).

Another controversial application of meta-analysis involved a test of the trophic cascade hypothesis (Brett and Goldman, 1996, 1997) and led to various papers criticizing, debating and elaborating on that meta-analysis (e.g. Osenberg *et al.*, 1999b). The meta-analyses of Brett and Goldman (1996, 1997) failed to support either of the two current models relating predation rate to prey numbers. Consequently, McCann *et al.* (1998) developed a new type of prey-dependent model, incorporating zooplankton interference and corresponding to a more reticulate food web, that more closely described the meta-analysis results. It is instructive to note the considerable overlap in the conceptual and methodological problems to be surmounted for the successful adoption of meta-analysis by the ecological and evolutionary communities. In evaluating the 'state-of-the-field' in meta-analysis in ecology, it is encouraging to note that we may be moving

beyond the integration of results for the purpose of testing existing hypotheses (e.g. Wand *et al.*, 1999), or providing parameter estimates for existing models (e.g. Medlyn *et al.*, 1999), to the development of new approaches or theories as a response to newly emerging perspectives resulting from meta-analytic synthesis.

VI. REAL AND PERCEIVED DIFFICULTIES IN ECOLOGICAL META-ANALYSIS

Although the utility of meta-analysis clearly is becoming more widely appreciated among ecologists, and statistical tools tailored to the needs of these researchers are being developed and disseminated, there remain a number of issues that present various impediments to their use of meta-analysis. While these issues are not unique to ecology and represent general classes of problems that apply in other disciplines as well, there is a distinct ecological aspect to their manifestation that we focus on here. In the following section we discuss difficulties arising from missing data, publication bias, data exclusion, non-independence among observations, and the combination of dissimilar data sets.

A. Missing Data

In our experience, the most serious impediment to the use of standard parametric meta-analysis methods for integrating ecological data is incomplete reporting of measurement statistics in the primary literature. Most common is the failure to unambiguously report sample sizes and/or standard deviations of response measures (i.e. papers commonly fail to report any measure of variation around the means, correlation coefficients, or whatever measure of outcome is being used to report the results of the experiment). We find this situation surprising given the attention paid to the need for statistical rigor in the editorial policies of most ecological journals, to say nothing of the curricula of our graduate education programs in ecology and evolution. Nonetheless, the net result is often a significant reduction in the number of usable studies and consequent loss of information in the analysis.

Several options are available to correct this situation. The long-term solution must be to raise publication standards. Sloppy reporting of results (such as the omission of sample sizes) should simply not be acceptable in mainstream ecological journals, even though it is currently fairly ubiquitous. Publication of data in graphical form is perfectly acceptable, though, and meta-analysts routinely scan and digitize data from figures (e.g. Gurevitch and Hedges, 2001). Standards can be enforced through

tightened editorial policies and more attention to this problem in the peer review process. In the short term, or for already published work, the meta-analyst can contact authors directly with a request to supply the missing data. Although this approach can produce unexpected levels of assistance and perhaps even lead to new professional contacts, it is almost always very time consuming, and generally has a low level of return.

B. Possible Options when Most Available Data Are Reported Poorly

One is still left with an enormous quantity of published ecological information in which sample sizes, standard deviations, or even means may be missing. Many reviewers want to know if there is any way to summarize such data, arguing that some information, even if of relatively low quality, is much better than no information. Clearly, if one can calculate a consistent measure of effect size across studies, one can estimate the (unweighted) grand mean. But how can one assess the magnitude of the variation around that mean, or test whether classes of studies differ in their mean effect sizes, since all of the relevant tests rely upon knowing the sampling variances?

One possibility that has been suggested is to use randomization methods (Adams et al., 1997). In this approach, confidence limits are constructed by bootstrapping the (unweighted) effect size data, and homogeneity tests are carried out using randomization, where the significance of the homogeneity statistic on the actual data is tested against a distribution created by randomly reassigning effects among categories many times. Continuous models can also be analyzed using randomization procedures on unweighted data to test for the significance of the slope (Rosenberg et al., 2000). The power and reliability of this approach will depend upon the error structure of the data set, yielding results that may be quite similar or quite different from those that would have been obtained if a parametric, weighted analysis had been possible (Gurevitch and Hedges, 1999).

Alternatively, one could carry out parametric tests such as ANOVA and regression on the unweighted effect sizes, but, as discussed above, the precision, statistical power and Type II error rates will be compromised, and the reliability of the results will likewise be unpredictable. A major dilemma with either of these approaches is not only that the results will not be as good as those from standard weighted procedures, but also that it is impossible to determine just how bad or good they are. Other problems also exist when one is forced into these sorts of compromises, including loss of the ability to separate within-study sampling error from between-study variation in true effects (Gurevitch and Hedges, 1999).

## C. Publication Bias and the Validity of Quantitative Synthesis

Publication bias is the selective publication of articles with particular results, rejecting papers that do not conform to these results (Begg, 1994). Publication bias typically refers to the tendency for editors to reject articles that fail to demonstrate statistically significant results (regardless of the particular findings), but it may also include a selective bias in favor of publishing those studies that confirm conventional wisdom. In contrast to the negative bias that can occur when combining the results of different studies using vote-counts (diminishing the frequency or magnitude of the effect being studied), publication bias typically is a systematic positive bias, resulting in a distortion that artifactually magnifies the magnitude of the effect of interest. Thus, if failure to reject the null hypothesis leads to higher publication rejection rates (or even self-selection by authors, leading to lower submission rates), then the combined results of a group of studies will be upwardly biased (Begg and Berlin, 1988).

It is important to recognize that publication bias, if it exists, will not only threaten the validity of a meta-analysis of that group of studies, but will affect *any* attempt to synthesize the literature on a topic, or to generalize from it (L.V. Hedges, personal communication). The conclusions of narrative reviews, then, are as vulnerable to publication bias as is meta-analysis. Evaluating the results of even a single paper potentially subjects one to the effects of publication bias; if a paper demonstrating an effect were put into the context of five more papers that failed to demonstrate that effect, the interpretation of the results of the first paper would no doubt be different. If those papers that fail to demonstrate the effect are rejected for publication, that context is gone and the interpretation of the results of that single published paper could lead to biased conclusions.

While one may suspect that publication bias exists, its extent and the degree to which it influences literature reviews remains conjectural unless it is investigated further. There have been a number of formal studies of publication bias in other disciplines (e.g., see Rosenberg *et al.*, 2000 for a brief review), and three papers that address this issue in ecology and evolution (Csada *et al.*, 1996; Bauchau, 1997; Palmer, 1999). The results of these investigations have been mixed, with some studies showing the likely existence of publication bias, while others suggest that its effects may be small. Meta-analysts have developed a series of approaches for detecting and quantifying possible cases of publication bias, and while they have not yet been applied in ecological meta-analyses, they would be well worth exploring (e.g. Kleijnen and Knipschild, 1992; Begg, 1994). Both graphical and analytical approaches can be used for detecting publication bias and estimating its magnitude.

The most familiar of the graphical methods are the examination of 'funnel plots', which are scatterplots of effect size graphed against sample size (Light and Pillemer, 1984; Figure 4). In the absence of publication bias, a funnel plot is expected to show a great deal of scatter in the magnitude of the measured effect sizes when sample sizes are small, with the scatter diminishing as sample size increases. This is because sampling theory predicts greater variance around the 'true' effect size when sample sizes are small; the measured effect sizes 'funnel down' or converge to the true effect size as sample sizes become larger. Publication bias can result in a dearth of points around zero at small sample sizes, if small studies reporting no significant effects are systematically rejected for publication (i.e. few or no effect sizes close to zero in value are reported for small sample sizes). Publication bias can also result in 'missing' points for larger negative (or positive) effect sizes, particularly at small sample sizes, if studies with unexpected values for the effect are routinely rejected. While funnel plots can be very useful for data exploration, they must be interpreted with some caution, because publication bias is not the only factor that can influence their appearance. For example, publication bias can result in a correlation between sample size and effect size, particularly when the true effect is moderate, because a 'bite' has been taken out of the collection of points for small effect sizes at low sample sizes (Begg, 1994; Figure 5). But other factors can also create such a correlation: for example, if experimentalists use larger sample sizes when hoping to detect small effects, but can 'get away with' smaller sample sizes when expecting larger
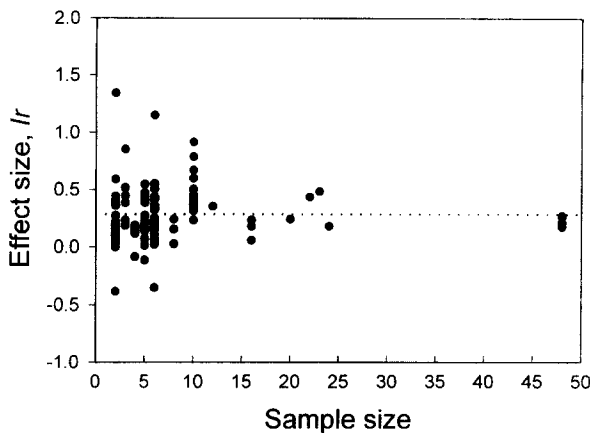
Fig. 4. 'Funnel plot' of effect size (log response ratio, lr) plotted against study sample size; data set as in Figure 2 for responses of trees to elevated CO_2. The dotted line indicates the mean effect size.

effects, the funnel plot will resemble Figure 5 as well. Funnel plots also have a number of other limitations (Wang and Bushman, 1998). A statistical test that evaluates the potential for publication bias in a data set in a similar manner to that of funnel plots is the weighted rank correlation test developed by Begg and Mazumdar (1994). So, although funnel plots are one of the most common tools for detecting publication bias, there are other alternatives that should be evaluated for that purpose.

Other graphical methods for data exploration in meta-analysis that are useful in detecting potential publication bias include weighted histograms and normal quantile plots (Wang and Bushman, 1998; Rosenberg *et al.*, 2000). Weighted histograms differ from ordinary histograms in that the bars are determined by the combined weights of the studies within a 'bin', rather than by frequencies, and they can be used to indicate various anomalies, including publication bias (see Rosenberg *et al.*, 2000). In a normal quantile plot, the data are ordered, and the quantiles are plotted against a standard normal distribution (i.e. with a mean of 0 and a standard deviation of 1.0). Confidence limits can be plotted around the results. Wang and Bushman (1998) detail three uses of such plots in
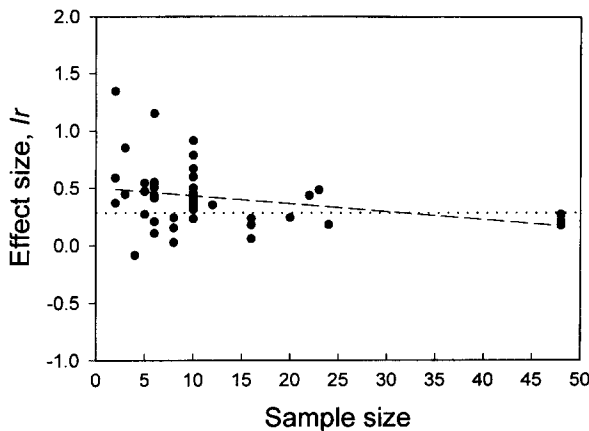
Fig. 5. Hypothetical funnel plot illustrating one possible indication of publication bias. Data set was created by manipulating the data set in Figure 4 by removing 56 of the studies in the lower left part of the graph of 102 total studies (i.e. removing many of those with the smallest effects at the smallest sample sizes) to create 'bias'. A regression line (dashed) indicates a negative relationship between effect size and sample size. Readers might note that it was necessary to remove more than half the published studies before such a negative relationship becomes visually apparent or statistically detectable.

meta-analysis: to check the normality assumption, to investigate whether all studies are likely to come from a single population of studies, and to search for publication bias. If the data fall relatively evenly along a straight line (with slope equal to the standard deviation and the Y coordinate of the center point equal to the mean), and within the confidence bands, the studies probably come from a single population of studies, are normally distributed, and probably do not evidence publication bias (Wang and Bushman, 1998). Publication bias and other causes of non-normal distributions might result in a gap in the points around zero, or in data that have peculiar non-linearities, such as a U-shaped distribution, with the tails (high and low values) extending upward outside of the confidence limits (Wang and Bushman, 1998; Figure 6). Graphical data exploration techniques are useful in meta-analysis for a variety of purposes, not limited to detecting publication bias (e.g. Rosenberg et al., 2000).

There are two general types of non-graphical statistical approaches to detecting and quantifying publication bias. The first is to estimate the magnitude of the problem by calculating a so-called 'fail-safe' number; the second is to model the bias and its impact on the results of the meta-analysis. Rosenthal (1979), imagining that studies that failed to reject the null hypothesis might remain unpublished in a filing cabinet, came up with the following solution to what he called the 'file drawer problem'. He proposed a simple method for estimating the number of unpublished studies with an effect of zero (the 'fail-safe number') that would be needed
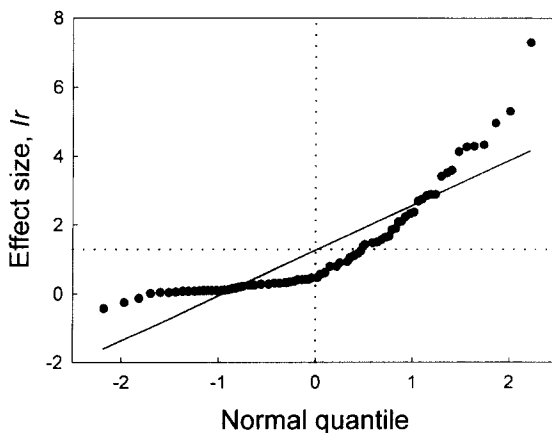
Fig. 6. Normal quantile plot (see text) from data set in Figure 4.

in order to overturn the conclusion that the average effect across studies was significantly greater (or less) than zero (Begg, 1994). If this number is 'sufficiently large', the reviewer might conclude that it is unlikely that so many unpublished studies actually exist, giving more confidence in the conclusion that there was a genuine effect when averaged across studies. 'Sufficiently large' is subjective, but a rule-of-thumb suggested by Rosenthal (1979) is $5k + 10$, where k is the number of studies. An approach for calculating 'fail-safe' number that is more similar to conventional weighted methods for analyzing data in meta-analysis has been suggested, as have other approaches (M.S. Rosenberg, personal communication, and see Rosenberg *et al.*, 2000). The second set of approaches to accounting for publication bias is to adjust the results of the analysis using weighted distribution theory (Begg, 1994). These complex methods are still under development and have not yet been widely applied in any field, including ecology.

D. Data Exclusion

When a meta-analysis is based upon published data, the question arises, how comprehensive does a literature search have to be? There will almost always be a need to restrict one's literature search in some way, if only in acknowledging financial and time constraints on the meta-analysis enterprise. Laird (1990) described attempting an exhaustive search for all published and unpublished literature on a topic to be 'one of life's mistakes that we only make once' (cf. Sharpe, 1997). However, there are two important points to be considered when devising ways to limit the scope of one's search. The first is the need to avoid biasing the search. Englund *et al.* (1999) discussed this issue specifically for ecological meta-analysis, recommending that several meta-analyses be performed using different selection criteria to test for the robustness of the conclusions. The meta-analyst should state explicitly, and *a priori*, the domain of his or her search (e.g. across certain journals, or years, or only using certain databases) and then attempt as thorough an initial search (i.e. to identify potentially useful papers) as possible within that domain. The second point, and arguably the more important one, relates to the conclusions one hopes to draw from the meta-analysis. If the goal is to generalize the meta-analysis findings to all studies that have been conducted in the field, regardless of publication status, time of publication or publication outlet, then a complete search, including unpublished results, would be preferred; if particular areas of the literature are systematically omitted, then one cannot extend one's conclusions to those areas (Sharpe, 1997). If the goal is to produce a quantitative estimate of central tendency across a

body of work, and perhaps explore the effect of various moderator variables on the magnitude of a treatment effect, a less intensive search may suffice. Of course, one must take care to select publications in an unbiased manner.

After potentially useful studies have been identified, further decisions must be made regarding which studies will actually be included in the meta-analysis. A persistent concern in meta-analysis is whether to combine high- and low-quality studies (Peipert and Bracken, 1997; Sharpe, 1997). Critics of meta-analysis have sometimes dismissed research syntheses as 'garbage in, garbage out'. Meta-analysts have often favored inclusion of all available studies, coding for quality so that the effects of including lower-quality studies could be tested directly (see below), while others favor screening for quality initially, excluding deficient studies from the data-base. Publication in peer-reviewed journals may in itself be viewed as adequate quality control, but failure to publish in such journals may be due to a number of factors, perceived quality being only one of them. Moreover, the selectivity or rigor of review varies considerably among journals. Neither approach is entirely satisfactory.

Rather than excluding results based on subjective assessments of quality, coding studies based on one or more measures of internal or statistical validity may be useful (Wortman, 1994). For example, criteria that are used to categorize the quality of each study might be whether treatments were fully replicated (versus pseudo-replicated), whether treatments were randomly assigned to subjects, whether there was confounding of factors, etc. With such 'research quality' moderator variables in place it is possible to test the hypothesis that results from 'poor-quality' studies differ from those of 'high-quality' research, and if so, should be excluded. Sharpe (1997) describes four separate meta-analyses in psychology that used an assortment of composite indexes of research quality that included measures of internal and statistical validity as well as other assessments. In no case, however, were categories defined on the basis of these indexes of research quality related to the magnitude of the treatment effect sizes. In other cases there has been no difference in mean results between high- and low-quality studies, but the latter exhibit more 'noise'. It is clear that there is no simple solution to this problem. However, insofar as 'study quality' is related to sample size or degree of replication, and hence precision, a standard weighted meta-analysis effect size estimate accounts for this source of variation among studies. That is, while we might expect poorly or pseudo-replicated studies to report spuriously high (or low) differences between treatment and control groups, we also expect them to have larger standard errors (i.e. lower precision), which will tend to discount the magnitude of between-group differences.

E. Non-Independence among Observations

One assumption of conventional meta-analysis, in common with most other inferential statistics, is that results of different individual studies, or observations within the database, are independent of one another. It is not known how robust or sensitive meta-analysis results are to different kinds of violations of this assumption, and, as in primary analyses, non-independence is still an unresolved problem. In primary research, control over experimental design can be employed to avoid violating assumptions of independence, and sometimes there are ways of statistically modeling and accounting for autocorrelation if it exists in a data set. In meta-analysis, the investigator has both a lack of control over experimental design and a more limited array of statistical tools at his or her disposal. For example, there is currently no way to accommodate non-independence of measurements through time, as one can in primary research using repeated measures ANOVA. However, if the between-measure correlations are known, multivariate methods have been developed to account for these dependencies (Hedges and Olkin, 1985). The issue of non-independence in ecological meta-analysis also is discussed by Gurevitch and Hedges (1999).

One approach to this issue has been to allow only one effect size estimate from any one published report (e.g. VanderWerf, 1992). Results appearing in separate publications are assumed to come from different experiments (and thus be independent of one another) and, conversely, results appearing in the same publication are considered necessarily statistically non-independent. However, both of these assumptions may be incorrect under some circumstances. Most obviously, those who choose to publish results of large, multi-parted experiments in single, large papers ('lumpers') will be treated differently from those who divide their results into several smaller papers ('splitters'). It also is not uncommon to encounter the same experimental results in more than one publication. Clearly, the 'publication' is not a quantity of sufficiently rigorous definition to be useful as a measure of statistical independence.

Fully crossed, multi-factorial experimental designs are common in ecological research, and present another dilemma. Where main treatment effects are independent in the context of the primary study, should they be considered independent in the context of the meta-analysis? For example, in studies of the effects of elevated CO_2 on plant growth, the CO_2 treatment is often crossed with another factor, such as soil nitrogen availability (e.g. Bazzaz and Miao, 1993). Authors typically report response means for all treatment combinations, such that in this example two CO_2 effect sizes can be calculated and added to the meta-analysis database: one CO_2 effect size for high nitrogen conditions (ambient CO_2,

high nitrogen compared to elevated CO_2, high nitrogen) and one CO_2 effect size for low nitrogen conditions (ambient CO_2, low nitrogen compared to elevated CO_2, low nitrogen). Although these results are from the same experiment and would normally be reported in the same publication, they might be argued to represent two independent estimates of the effect of CO_2 on plant growth. Does it make sense to combine them? The answer depends upon the goals of the summary. Information about responses over a range of nitrogen availability may be sufficiently valuable that one would not want to omit either response, because including them both enables one to generalize more broadly. On the other hand, the details of how the experiment was conducted might conceivably cause the responses to elevated CO_2 to be so highly correlated at the two nitrogen levels that they do not provide independent information about those responses, compromising their value for the meta-analysis. The decision to include or omit studies is a scientific as well as a statistical one.

The situation may be made more complicated, however, by the incorporation of multiple species, or genotypes, within a factorial design. In the above example, each experimental unit consisting of a specific CO_2 and soil nitrogen treatment could contain two (or more) different species, resulting in a factorial, split-plot design, with CO_2 and nitrogen as the main plots and species as the sub-plot. In this example, four 'independent' CO_2 effect size estimates could be calculated (response to elevated CO_2 at low N and at high N for each of the two species). Although the species share an experimental unit, and are therefore not strictly independent in their response to the treatments, they could be so different biologically (e.g. wheat versus oak) that we would reasonably expect them to respond very differently to the treatments. Given a lack of strict statistical independence among effect size estimates in such an experiment, one could argue that species differences should be ignored and only the average effect size across species be included in the meta-analysis, or that the results from only one of the species be used. Because this approach can result in a considerable loss of information, an alternative is to consider the responses of different species to a common treatment within a single study as effectively independent of one another. In the latter case, a single study examining many species can contribute a large number of observations to a meta-analysis.

There may be other, less obvious, sources of non-independence among effect size estimates in an ecological meta-analysis database. The magnitude or direction of treatment effect sizes among species could be correlated with their degree of phylogenetic relatedness. The problem of phylogenetic non-independence has been widely discussed in the evolutionary literature (Harvey and Pagel, 1991; Silvertown and Dodd,

1996) and it is likely that some of the same issues apply to meta-analysis. However, while it may be possible for a primary researcher to select species based on known phylogenetic relationships, this is rarely the case when synthesizing research results. It may be impossible in most cases to conduct phylogenetically independent contrasts *sensu stricto* (Felsenstein, 1985) among effect size estimates (P.S. Curtis, unpublished data). A variation of non-independence based on evolutionary phylogeny is that of non-independence based on academic phylogeny. That is, it might be hypothesized that researchers working in the same lab, or simply trained in the same lab, might, by virtue of common techniques or viewpoints, publish results that are more similar in terms of treatment effects sizes than do those from unrelated labs. A preliminary analysis of patterns of results from the elevated CO_2 literature revealed no such relationship (L. Hedges, personal communication) but this issue has certainly not been widely investigated. Research reviewers should understand that when independence of results within the meta-analysis database is not achieved, the standard error of mean effects will be underestimated, leading to increased probability of Type I error (Gleser and Olkin, 1994). A prudent course of action under these circumstances might be to use a more conservative standard for rejecting the null hypothesis in significance tests (i.e. the α level) from 0.05 to 0.01 or even lower.

F. Apples, Oranges and Other Challenges to the Validity of Meta-analysis

Another objection raised by critics of meta-analysis is that these techniques combine unlike studies, producing meaningless results. Glass, an early developer and proponent of meta-analysis in psychology, replied to this contention that it is a good thing to combine apples and oranges if one wishes to understand something about fruit (as cited by Smith *et al.*, 1980). This argument has, however, not satisfied all critics. Clearly, each meta-analysis must provide a justification for the scope of the review. Too broad a focus will risk meaningless results, while too narrow a focus will not only risk being uninteresting, but can also encourage misleading extrapolation if the conclusions are extended to studies outside the domain of the synthesis (Sharpe, 1977). Evaluating results for agreement among studies' outcomes using homogeneity tests, and grouping similar studies together, can also be a way to address this issue. Unfortunately, the number of 'similar' studies (however that is defined) available to partition by study category in a given meta-analysis may not be great enough to provide for meaningful subdivision. For example, if

the data base contains three studies on apples, two on oranges, and one each on bananas, pomegranates and mangoes, meaningful partitioning into homogeneous categories based on kinds of fruit may be difficult. If all of the studies agree in their responses, however, this may be neither necessary or desirable. Before combining effect sizes, Hedges *et al.* (1999) recommend comparison of the between-study variance to the average within-study variance. If the between-study variation is many times the average within-study sampling error variation, the studies may be too dissimilar to warrant combination. Ultimately, the question of 'apples and oranges' raises fundamental issues regarding the nature of generalizing from experimental results to truths about nature that cannot be resolved simply.

In addition to discussing the threats to meta-analysis validity addressed above (publication bias, 'garbage in garbage out' and 'apples and oranges'), Sharpe (1997) also considered other possible explanations for continued opposition to the application of meta-analysis among some in the social sciences, despite its general acceptance. Poor first impressions based upon early meta-analyses, condemning meta-analysis on the basis of problematic individual applications of this approach, assuming that the results of meta-analysis will appear to be more definitive than they actually are and cut off future research, and the purportedly descriptive or atheoretical nature of quantitative synthesis are all possible reasons. We would add that meta-analysis is by its very nature a powerful but fairly blunt instrument, and that expecting it to reveal fineness and subtlety of detail risks disappointment. Quantitative research synthesis is in its infancy, and we are still exploring what it can do well, and what tasks it inevitably will do poorly.

VII. CASE STUDIES OF RECENT APPLICATIONS OF META-ANALYSIS IN GLOBAL CHANGE BIOLOGY

Research syntheses can be organized in a number of ways depending on data availability and study objectives. In this section we describe three different approaches to using meta-analysis as a data integration tool as applied by three different research groups examining effects of global change on plants and ecosystems. The first example illustrates the approach of comprehensive literature searching, in which previously published studies are mined for any data relevant to the review objectives. The second example is one where a group of researchers combined data from their independent and otherwise uncoordinated

experiments for purposes of hypothesis testing and model development. The final example describes the efforts of a large group of independent investigators who agreed to collect data in a common format with the goal of conducting coordinated integration and synthesis of data from their individual studies.

A. The Elevated CO_2 Meta-analysis Project (CO_2MAP)

Understanding the ecological consequences of rising atmospheric CO_2 has engaged scientists in a variety of disciplines for over 20 years. Strain and Cure (1994) documented an average of 90 elevated CO_2 papers per year published between 1989 and 1992, while a recent search of the Institute for Scientific Information database for a single year, 1998, yielded over 500 elevated CO_2 citations. Synthesis of this body of research resulted in an average of 19 review articles published per year between 1989 and 1992 and (at least) 12 reviews published in 1998. CO_2MAP was initiated in 1995 at the Ohio State University with three primary objectives: (1) the establishment of a comprehensive, digital database of all studies published to date on the response of vegetation to elevated CO_2; (2) integration and synthesis of these results using meta-analytic methods; and (3) distribution of this database via a server at the Carbon Dioxide Information Analysis Center of Oak Ridge National Laboratory. This approach typifies that of many meta-analyses in which a small research group focuses solely on previously published material, but the cumulative, ongoing database is more like the approach taken specifically in some areas of medical meta-analysis.

The CO_2MAP database was originally restricted to woody plant species and covered 83 experiments with 41 different species. The literature search strategies and data extraction methods used are described in more detail below (see section VIII). Curtis (1996), in the first meta-analysis of the elevated CO_2 literature, provided quantitative support for several general conclusions reached by previous reviewers and contributed a number of additional insights into the nature of CO_2 responses in trees. These results were extended by Curtis and Wang (1998), following the addition of 415 studies and 18 species to the database. They also introduced the use of the weighted log ratio as an estimator of CO_2 effect size. This metric was further described by Hedges *et al.* (1999). The CO_2MAP database was expanded to include herbaceous species in 1997. Using this database, Wand *et al.* (1999) examined the responses of C_3 and C_4 species to elevated CO_2 as a test of differences among these plant functional types. The complete meta-database for both woody and

herbaceous species is available at: http://cdiac.esd.ornl.gov (Curtis *et al.*, 1999; Jones *et al.*, 1999).

B. CO$_2$ Model/Experiments Activity for Improved Links (CMEAL)

This multi-investigator project was initiated in 1995 with a broad objective to re-evaluate the way elevated CO$_2$ responses are handled in ecosystem models (LeFlohic, 1998). In addition to the detailed intercomparison of a number of current ecosystem-level models (e.g. BIOME-BGC, CENTURY, GEM) an important part of this project was a meta-analysis of existing data for the purpose of model parameterization, setting constraints on model structure, and formulating tests of alternative hypotheses regarding ecosystem responses to CO$_2$ enrichment. The meta-analysis component of this work was also envisioned as a means to search for mechanisms controlling ecosystem carbon storage that could be further explored both experimentally and with models. Thus, the quantitative integration of existing data was seen as a fundamental step in the linkage of empirical data to models and back to the generation of new experimental results.

Unlike CO$_2$MAP, CMEAL did not draw from the entire elevated CO$_2$ literature but rather focused on a relatively small number of experiments (~20) conducted under more natural conditions, whose results were regarded *a priori* as being more relevant to ecosystem-scale questions than, for example, results from experiments using potted plants grown inside controlled environment chambers. Scientists representing six of these experiments were co-investigators within CMEAL and agreed to contribute primary data to the meta-analysis. These six experiments were not, however, otherwise coordinating data collection or analysis efforts. That is, while the targeted experiments were similar with regard to certain aspects of scale and methodology, there was no necessary standardization of measurement parameters or techniques.

Peterson *et al.* (1999) conducted a meta-analysis of the responses of 39 plant species grown at ambient CO$_2$ and 10 species grown at both ambient and elevated CO$_2$. They used a regression model to evaluate the effects of elevated CO$_2$ on the photosynthesis:leaf nitrogen relationship, finding that CO$_2$ enrichment significantly increased photosynthetic nitrogen use efficiency, but that the magnitude of this effect depended on the vegetation type (e.g. evergreen versus deciduous trees). Although this aspect of the meta-analysis was quite successful, the lack of standardization among investigators in data collection and reporting procedures presented considerable difficulties (A. Peterson, personal

communication). Efforts to integrate the meta-analysis results into model development are ongoing.

C. The International Tundra Experiment (ITEX)

The International Tundra Experiment (ITEX), established in 1990, is a collaborative research effort examining the response of circumpolar plant species to passive increases in summer warming (Henry and Molau, 1997). There are nearly 40 ITEX sites located in arctic and alpine habitats in 13 countries. ITEX was designed to provide an understanding of how the same taxa located in different geographical areas respond to atmospheric warming and natural climatic variability, and how changes in those species and their environments will affect overall ecosystem structure and function. A manual was developed by ITEX participants defining common experimental protocols, including field chamber design, study species and data collection. However, experimental methods differed slightly among investigators, and experiments were initiated at different times among the many locations. In this sense, ITEX is not a single large experiment, but rather a collection of closely coordinated studies, making meta-analysis the appropriate method for statistical analysis of the entire ITEX dataset.

In 1996 a workshop was held at the National Center for Ecological Analysis and Synthesis to conduct a meta-analysis of ITEX data collected during the first years of the experiment. Prior to and during the workshop, individual participants from around the globe organized their data into a final, standardized format. Several teams then conducted exploratory meta-analyses of different response variables. The results of the ITEX meta-analysis – covering the first four years of this collaborative project – were published recently in *Ecological Monographs* (Arft *et al.*, 1999). The meta-analysis showed that tundra plants in general increased vegetative and reproductive growth in response to warmer summer temperatures. Vegetative biomass increased most in low arctic and alpine species, while reproduction increased most in high arctic plants. Herbaceous species tended to respond more strongly than woody species, particularly in vegetative growth.

Using meta-analysis to synthesize the results of the coordinated ITEX data proved highly successful, but was not without its difficulties. In particular, gathering and collating many disparate data sets into one common data format ready for analysis was a considerable task, requiring substantially more effort than anticipated. The strategy of having multiple investigative teams conduct meta-analyses of different data components was useful for exploratory purposes during the workshop, but was not adequate for the final synthesis (due perhaps to the limited time

frame available for the group to work together on the analysis, as well as to a possible case of 'too many cooks' going in too many different directions). In the end, a single analyst working closely with a few investigators was most successful in integrating the results to complete the research synthesis. The resulting manuscript was then circulated to the main group for comments, revised, approved by the group, and finally published.

VIII. PRACTICAL CONSIDERATIONS IN CARRYING OUT META-ANALYSES IN ECOLOGY

In this section we provide an overview of how the CO_2 Meta-Analysis Project operates, as a practical reference for other investigators. As described above, CO_2MAP is fairly typical for a large meta-analysis research program, relying on the published literature for all input data. The project is organized around two separate but integrated activities: (a) identifying and acquiring relevant literature, and (b) using specific hypotheses to frame meta-analyses of that literature. While it may be true that attempting an exhaustive search for all published and unpublished literature on a subject is a once-in-a-lifetime mistake, our experience has shown that meta-analyses involving large numbers of studies are certainly possible, but also clearly benefit from good organization and advanced planning.

A. Data Identification

Data identification at CO_2MAP involves literature searches, manuscript acquisition and archiving, and evaluation of manuscripts for relevance to particular research questions (Figure 7). Identification of elevated CO_2 papers relies primarily on on-line bibliographic services such as those provided by the Institute for Scientific Information. This phase of the search is as comprehensive as possible, although the focus is strictly on studies published in the peer-reviewed literature. That is, the goal is to attempt to maintain a record of all elevated CO_2 studies published. Once identified, a reference is entered into the master electronic bibliography and assigned an accession number. This number identifies not only that specific citation, but also the paper copy of the publication if it is acquired, and any data collected from that publication. Abstracts and keywords are included in the bibliographic database whenever possible, facilitating

Data Identification

External search
  • Initial manuscript evaluation (abstract, keywords)
Enter citation to **BIBLIOGRAPHIC DATABASE**
Internal search
  • Initial manuscript evaluation (abstract, keywords)
Manuscript acquisition, archive
In-hand evaluation of manuscript:
                is it appropriate for analysis?

External search:
  • electronic database with exportable citations, key words, abstracts

Bibliographic database:
  • searchable electronic database with import/export capabilities
  • include citations, key words, abstracts

Maintain clean and annotated copies of mss.

Data Collection

Data extraction, recording
  • **META DATA**: one record per study.
  • **RESPONSE DATA**: multiple records per study.
  • Hardcopy data sheets.
Data Entry
  • Use of templates with error checking to enter data to spreadsheet.
Data Checking
  • Comparision of records in spreadsheet to hardcopy data sheets.
Data Storage
  • **MANUSCRIPTS** together with **HARDCOPY DATA SHEETS** in file.
  • Spreadsheets: original, archival copy plus subsequent manipulated datasets.
  • Electronic back up.

Data extraction:
  • direct from text and tables
  • indirect from digitized figures, quantification software
Data recording:
  • data recorded on standardized hardcopy data sheets
  • electronic, printable versions facilitate data recording process

Data entry:
  • data entry templates with error check controls can help reduce errors, particularly with inexperienced assistants
  • data entry templates on screen appear identical to hardcopy data sheets

Maintain hardcopy and electronic archives.
Electronic back up essential.

Meta-Analysis

Incremental preparation of datasets.

Retain dataset integrity.

Fig. 7. Schematic overview of the CO_2 Meta-Analysis Project procedures.

future searches of the database. Copies of papers are acquired for reference purposes and evaluation.

The actual meta-analyses begin with a search of the master bibliographic database, acquisition of potentially relevant papers, and evaluation of each paper for its usefulness based upon the constraints of the specific hypotheses or analyses being conducted. Two copies of each paper are retained; an archival clean copy, and a second copy for detailed annotations on the specifics of data collection, and for later data checking and quality control.

B. Data Collection

Data collection involves four components; data extraction and recording, data entry, data checking and data storage (Figure 7). Two types of data are extracted from each study: meta-data (methods and experimental conditions) and response data. Meta-data are used to provide background information for subsequent analyses or interpretation and to place studies within specified categorical groups. Since meta-data often include methodological details that will not be used as categorical variables, a separate electronic file is maintained of all meta-data, in which a single record is identified with an individual study's accession number.

Response data are the numerical data quantifying the species or system responses to the experimental treatments. Response data form the core of the data record although, as noted above, each data record also contains the meta-data necessary to partition responses by categorical groups during the actual meta-analysis. Data records are identified by their accession number and by an additional page number. That is, each data record has an accession number relating it to the paper from which it came, and a unique page number identifying it as one of the (perhaps many) records to have been derived from that paper. This dual identification system is critical for accuracy in data entry and ease of data checking.

Response data are reported in the primary literature as text, tables or figures. To extract data from figures, we first scan and digitize the figures and then use image analysis software to determine values (typically means and standard errors) from the digitized information (e.g. see Gurevitch and Hedges, 2001). Meta-data and response data are recorded on standardized data sheets, with one data record per sheet. These data may be recorded entirely by hand, but multiple copies of partially completed data sheets may help expedite data collection. For example, the species name or CO_2 treatment levels may be the same for all records, and can be pre-printed.

Spreadsheet templates are used for data entry into the computer. These look identical to the paper data sheet, helping to minimize transposing errors during data entry. Internal error checking is performed by the spreadsheet software as each cell is completed. Other software specifically designed for expediting data entry and error checking is also available commercially. Post-entry data checking is typically performed by someone other than the person conducing the data entry. Data storage includes electronic backups, as well as the annotated copy of each manuscript from which data are mined.

These methods for searching and mining the literature and for data entry work well for the CO_2 Meta-Analysis Project, and may be useful as a framework for other projects. While the CO_2MAP work is on-going, a

similar approach – and identical concerns – would apply for a single literature-based meta-analysis.

IX. CONCLUSIONS: NEW DIRECTIONS

Meta-analysis is poised to begin having a substantial impact on ecological research and on ecological thinking. It will almost certainly eventually be adopted for use in informing policy, as it has been in other disciplines where its use in basic research has become fairly standard. Some ecologists and evolutionists will never accept it, but a great many will surely embrace it because the tools it offers are too powerful to ignore. As with any set of statistical techniques, there is a risk that meta-analysis will be used poorly (as, for example, ANOVA has been extensively misused in ecology). One has to understand the methods one is using, including their limitations, assumptions and pitfalls, whether one is measuring CO_2 uptake, doing a t-test, or carrying out a meta-analysis. Thoughtful implementation of these methods will surely provide a wealth of information that is currently inaccessible to us; inappropriate application of meta-analysis will surely produce misleading results.

Many developments can be made in the usefulness and accuracy of meta-analysis in ecology, as well as in other disciplines. These range from the purely statistical (e.g. extension of GLM (General Linear Models) approaches in meta-analysis; Rosenberg *et al.*, 2000) to those concerned largely with ecological and evolutionary issues, to those somewhere in between. As an example of the latter, a better understanding of how robust meta-analysis results are to various types of non-independence (as discussed above) would be enormously useful (as it would be for primary analyses). Not only do better methods need to be worked out to detect and correct for publication bias, but in a larger context, greater awareness of publication bias may change the criteria for publication. As these issues continue to be discussed, reviewers and editors may begin to realize that results that are 'not significant' are not necessarily 'not informative' or 'not interesting'. Wider acceptance and implementation of meta-analysis may therefore change the way papers are judged for acceptance for publication. As anyone who has been involved in this enterprise would agree, we fervently hope that this will also include higher standards for data reporting (i.e. explaining clearly what was done in an experiment, reporting sample sizes and standard deviations). As meta-analysis becomes more widely known in ecology, people will also be interested in applying these techniques to kinds of data that require new methodology, for example, to non-experimental data.

We end this review as we began it, by predicting that the incorporation of meta-analysis as a routine and familiar approach in ecology and evolution will fundamentally, if subtly, change the nature of these scientific disciplines in both predictable and unpredictable ways. If implemented thoughtfully, the change should be for the better.

ACKNOWLEDGMENTS

REFERENCES

Abouheif, E. and Fairbairn, D.J. (1997). A comparative analysis of allometry for sexual size dimorphism: assessing Rensch's rule. *Am. Nat.* **149**, 540–562.

Adams, D.C., Gurevitch, J. and Rosenberg, M.S. (1997). Resampling tests for meta-analysis of ecological data. *Ecology* **78**, 1277–1283.

Arft, A.M., Walker, M.D., Gurevitch, J., Alatalo, J.M., Bret-Harte, M.S., Dale, M., Diemer, M., Gugerli, F., Henry, G.H.R., Jones, M.H., Hollister, R.D., Jonsdottir, I.S., Laine, K., Levesque, E., Marion, G.M., Molau, U., Nordenhall, U., Razzhivin, V., Robinson, C.H., Starr, G., Stenstrom, A., Stenstrom, M., Totland, O., Turner, P., Walker, L.J., Webber, P.J., Welker, J.M. and Wookey, P.A. (1999). Responses of tundra plants to experimental warming: meta-analysis of the International Tundra Experiment. *Ecol. Monogr.* **69**, 491–511.

Arnqvist, G. and Nilsson, T. (2000). The evolution of polyandry: multiple mating and female fitness in insects. *Anim. Behav.* **60**, 145–164.

Arnqvist, G. and Wooster, D. (1995a). Meta-analysis – synthesizing research findings in ecology and evolution. *TREE* **10**, 236–240.

Arnqvist, G. and Wooster, D. (1995b). Statistical power of methods of meta-analysis – reply. *TREE* **10**, 460–461.

Arnqvist, G., Rowe, L., Krupa, J.J. and Sih, A. (1996). Assortative mating by size: a meta-analysis of mating patterns in water striders. *Evol. Ecol.* **10**, 265–284.

Bauchau, V. (1997). Is there a 'file drawer problem' in biological research? *Oikos* **79**, 407–409.

Bazzaz, F.A. and Miao, S.L. (1993). Successional status, seed size, and responses of tree seedlings to CO_2, light and nutrients. *Ecology* **74**, 104–112.

Begg, C.B. (1994). Publication bias. In: *The Handbook of Research Synthesis* (Ed. by H. Cooper and L.V. Hedges), pp. 399–410. Russell Sage Foundation, New York.

Begg, C.B. and Berlin, J.A. (1988). Publication bias: A problem in interpreting medical data. *J. Roy. Stat. Soc. Ser. A*. **151**, 419–463.

Begg, C.B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50**, 1088–1101.

Bender, D.J., Contreras, T.A. and Fahrig, L. (1998). Habitat loss and population decline: a meta-analysis of the patch size effect. *Ecology* **79**, 517–533.

Blanckenhorn, W.U., Reusch, T. and Muhlhauser, C. (1998). Fluctuating asymmetry, body size and sexual selection in the dung fly *Sepsis cynipsea* – testing the good genes – assumptions and predictions. *J. Evol. Biol.* **11**, 735–753.

Brett, M.T. (1997). Meta-analysis in ecology. *Bull. Ecol. Soc. Am.* **78**, 92–94.

Brett, M.T. and Goldman, C.R. (1996). A meta-analysis of the freshwater trophic cascade. *Proc. Natl Acad. Sci. USA* **93**, 7723–7726.

Brett, M.T. and Goldman, C.R. (1997). Consumer versus resource control in freshwater pelagic food webs. *Science* **275**, 384–386.

Brett, M.T. and Muller-Navarra, D.C. (1997). The role of highly unsaturated fatty acids in aquatic food web processes. *Freshw. Biol.* **38**, 483–499.

Britten, H.B. (1996). Meta-analyses of the association between multilocus heterozygosity and fitness. *Evolution* **50**, 2158–2164.

Brook, B.W., O'Grady, J.J., Chapman, A.P., Burgman, M.A., Akcakaya, H.R. and Frankham, R. (2000). Predictive accuracy of population viability analysis in conservation biology. *Nature* **404**, 385–387.

Byers, D.L. and Waller, D.M. (1999). Do plant populations purge their genetic load? Effects of population size and mating history on inbreeding depression. *Ann. Rev. Ecol. Syst.* **30**, 479–513.

Cebrian, J., Duarte, C.M., Agawin, N.S.R. and Merino, M. (1998). Leaf growth response to simulated herbivory: a comparison among seagrass species. *J. Exp. Mar. Biol. Ecol.* **220**, 67–81.

Chalmers, I., Enkin, M. and Keirse, M.J.N.C. (1989). *Effective Care in Pregnancy and Childbirth*. Oxford University Press, Oxford.

Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *J. Roy. Stat. Soc.* **4** (suppl.), 102–118.

Collie, J.S., Hall, S.J., Kaiser, M.J. and Poiner, I.R. (2000). A quantitative analysis of fishing impacts on shelf-sea benthos. *J. Anim. Ecol.* **69**, 785–798.

Connor, E.F., Courtney, A.C. and Yoder, J.M. (2000). Individuals–area relationships: The relationship between animal population density and area. *Ecology* **81**, 734–748.

Cooper, H. and Dorr, N. (1995). Race comparisons on need for achievement: A meta-analytic alternative to Graham's narrative review. *Rev. Ed. Res.* **65**, 483–508.

Cooper, H. and Hedges, L.V. (Editors) (1994a). *The Handbook of Research Synthesis*. Russell Sage Foundation, New York.

Cooper, H. and Hedges, L.V. (1994b). Research synthesis as a scientific enterprise. In: *The Handbook of Research Synthesis* (Ed. by H. Cooper and L.V. Hedges), pp. 3–14. Russell Sage Foundation, New York.

Cote, I.M. and Sutherland, W.J. (1997). The effectiveness of removing predators to protect bird populations. *Cons. Biol.* **11**, 395–405.

Csada, R.D., James, P.C. and Espie, R.H.M. (1996). The 'file drawer problem' of non-significant results: does it apply to biological research? *Oikos* **76**, 591–593.

Curtis, P.S. (1996). A meta-analysis of leaf gas exchange and nitrogen in trees grown under elevated carbon dioxide. *Plant Cell Envir.* **19**, 127–137.

Curtis, P.S. and Wang, X.Z. (1998). A meta-analysis of elevated CO_2 effects on woody plant mass, form, and physiology. *Oecologia* **113**, 299–313.

Curtis, P.S., Cushman, R.M. and Brenkert, A.L. (1999). *A database of woody vegetation responses to elevated atmospheric CO_2*. ORNL/CDIAC-120, NDP-072. Carbon Dioxide Information Analysis Center, US Department of Energy, Oak Ridge National Laboratory, Oak Ridge, TN.

Dahl, J. and Greenberg, L. (1996). Impact on stream benthic prey by benthic vs. drift feeding predators: a meta-analysis. *Oikos* **77**, 177–181.

Davidoff, F. (1999). In the teeth of the evidence: The curious case of evidence-based medicine. *Mount Sinai J. Med.* **66**, 75–83.

Dodds, W.K. (1997). Interspecific interactions: constructing a general neutral model for interaction type. *Oikos* **78**, 377–383.

Dolman, P.M. and Sutherland, W.J. (1997). Spatial patterns of depletion imposed by foraging vertebrates: theory, review and meta-analysis. *J. Anim. Ecol.* **66**, 481–494.

Downing J.A., Osenberg, C.W. and Sarnelle, O. (1999). Meta-analysis of marine nutrient-enrichment experiments: variation in the magnitude of nutrient limitation. *Ecology* **80**, 1157–1167.

Ellison, A.M. (1996). An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecol. Appl.* **6**, 1036–1046.

Englund, G. and Evander, D. (1999). Interactions between sculpins, net-spinning caddis larvae and midge larvae. *Oikos* **85**, 117–126.

Englund, G., Sarnelle, O. and Cooper, S.D. (1999). The importance of data-selection criteria: meta-analyses of stream predation experiments. *Ecology* **80**, 1132–1141.

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15.

FernandezDuque, E. (1997). Comparing and combining data across studies: alternatives to significance testing. *Oikos* **79**, 616–618.

FernandezDuque, E. and Valeggia, C. (1994). Meta-analysis – a valuable tool in conservation research. *Cons. Biol.* **8**, 555–561.

Fisher, R.A. (1932). *Statistical Methods for Research Workers*, 4th edn. Oliver and Boyd, London.

Fiske, P., Rintamaki, P.T. and Karvonen, E. (1998). Mating success in lekking males: a meta-analysis. *Behav. Ecol.* **9**, 328–338.

Folt, C.L., Nislow, K.H. and Power, M.E. (1998). Implications of temporal and spatial scale for Atlantic salmon (*Salmo salar*) research. *Can. J. Fisheries Aquat. Sci.* **55**, 9–21.

Frankham, R. (1996). Relationship of genetic variation to population size in wildlife. *Cons. Biol.* **10**, 1500–1508.

Glass, G.V. (1976). Primary, secondary and meta-analysis. *Educational Researcher* **5**, 3–8.

Gleser, L.J. and Olkin, I. (1994). Stochastically dependent effect sizes. In: *The Handbook of Research Synthesis* (Ed. by H. Cooper and L.V. Hedges), pp. 339–356. Russell Sage Foundation, New York.

Gliwicz, J. and Glowacka, B. (2000). Differential responses of *Clethrionomys* species to forest disturbance in Europe and North America. *Can. J. Zool.* **78**, 1340–1348.

Goldberg, D.E., Rajaniemi, T., Gurevitch, J. and Stewart-Oaten, A. (1999). Empirical approaches to quantifying interaction intensity: competition and facilitation along productivity gradients. *Ecology* **80**, 1118–1131.

Gough, L., Osenberg, C.W., Gross, K.L. and Collins, S.L. (2000). Fertilization effects on species density and primary productivity in herbaceous plant communities. *Oikos* **89**, 428–439.

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

Gurevitch, J. and Hedges, L.V. (1999). Statistical issues in ecological meta-analysis. *Ecology* **80**, 1142–1149.

Gurevitch, J. and Hedges, L.V. (2001). Meta-analysis: combining the results of independent experiments. In: *Design and Analysis of Ecological Experiments,* 2nd edn. (Ed. by S.M. Scheiner and J. Gurevitch), pp. 347–369. Oxford University Press, New York and Oxford.

Gurevitch, J., Morrow, L.L., Wallace, A. and Walsh, J.S. (1992). A meta-analysis of competition in field experiments. *Am. Nat.* **140**, 539–572.

Gurevitch, J., Morrison, J.A. and Hedges, L.V. (2000). The interaction of competition and predation: a meta-analysis of field experiments. *Am. Natur.* **155**, 435–453.

Hamilton, W.J. and Poulin, R. (1997). The Hamilton and Zuk hypothesis revisited: a meta-analytical approach. *Behaviour* **134**, 299–320.

Hartley, M.J. and Hunter, M.L. (1998). A meta-analysis of forest cover, edge effects, and artificial nest predation rates. *Cons. Biol.* **12**, 465–469.

Harvey, P.H. and Pagel, M.D. (1991). *The Comparative Method in Evolutionary Biology.* Oxford University Press, Oxford.

Hechtel, L.J. and Juliano, S.A. (1997). Effects of a predator on prey metamorphosis: plastic responses by prey or selective mortality? *Ecology* **78**, 838–851.

Hedges, L.V. and Olkin, I. (1980). Vote counting methods in research synthesis. *Psychol. Bull.* **88**, 359–369.

Hedges, L.V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Academic Press, New York.

Hedges, L.V. and Olkin, I. (2001). *Statistical Methods for Meta-Analysis in the Medical and Social Sciences.* Academic Press, New York (in press).

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

Henry, G.H.R. and Molau, U. (1997). Tundra plants and climate change: the International Tundra Experiment (ITEX). *Global Change Biol.* **3** (Suppl. 1), 1–9.

Herrera, C.M., Jordano, P., Guitian, J. and Traveset, A. (1998). Annual variability in seed production by woody plants and the masting concept: reassessment of principles and relationship to pollination and seed dispersal. *Am. Nat.* **152**, 576–594.

Hilborn, R. and Liermann, M. (1998). Standing on the shoulders of giants: learning from experience in fisheries. *Rev. Fish Biol. Fisheries* **8**, 273–283.

Hollister, R.D. and Webber, P.J. (2000). Biotic validation of small open-top chambers in a tundra ecosystem. *Global Change Biol.* **6**, 835–842.

Houle, D. (1997). A meta-analysis of the heritability of developmental stability – comment. *J. Evol. Biol.* **10**, 17–20.

Hugueny, B. and Guegan, J.F. (1997). Community nestedness and the proper way to assess statistical significance by Monte-Carlo tests: comments. *Oikos* **80**, 572–574.

Hunt, M. (1997). *How Science Takes Stock: The Story of Meta-analysis*. Russell Sage Foundation, New York.

Hurlbert, S.H. (1994). Old shibboleths and new syntheses. *TREE* **9**, 495–496.

Jarvinen, A. (1991). A meta-analytic study of the effects of female age on laying-date and clutch-size in the great tit *Parus-major* and the pied flycatcher *Ficedula hypoleuca. Ibis* **133**, 62–66.

Jones, A. (2000). Effects of cattle grazing on North American arid ecosystems: a quantitative review. *West. N. Am. Nat.* **60**, 155–164.

Jones, M.H., Curtis, P.S., Cushman, R.M. and Brenkert, A.L. (1999). *A comprehensive database of herbaceous vegetation responses to elevated atmospheric CO_2.* ORNL/CDIAC-124, NDP-073. Carbon Dioxide Information Analysis Center, US Department of Energy, Oak Ridge National Laboratory, Oak Ridge, TN.

Kleijnen, J. and Knipschild, P. (1992). Review articles and publication bias. *Arzneimittel-Forschung/Drug Res.* **42**, 587–591.

Koricheva, J., Larsson, S. and Haukioja, E. (1998a). Insect performance on experimentally stressed woody plants: a meta-analysis. *Ann. Rev. Entomol.* **43**, 195–216.

Koricheva, J., Larsson, S., Haukioja, E. and Keinanen, M. (1998b). Regulation of woody plant secondary metabolism by resource availability: hypothesis testing by means of meta-analysis. *Oikos* **83**, 212–226.

Kraak, S.B.M. (1997). Fluctuating around directional asymmetry? *TREE* **12**, 230.

Laird, N.M. (1990). A discussion of the Aphasia Study. In: *The Future of Meta-analysis* (Ed. by K.W. Wachter and M.L. Straf), pp. 47–52. Russell Sage Foundation, New York.

Lardicci, C. and Rossi, F. (1998). Detection of stress on macrozoobenthos: evaluation of some methods in a coastal Mediterranean lagoon. *Mar. Environ. Res.* **45**, 367–386.

Leamy, L. (1997). Is developmental stability heritable? Commentary. *J. Evol. Biol.* **10**, 21–29.

LeFlohic, J. (1998). CMEAL. CO_2 Models/Experiments Activity for Improved Links (Online): *http://www.gcte-focus1.org/activities/Cmeal/cmeal.html* (2001, Jan. 4).

Lempa, K., Martel, J., Koricheva, J., Haukioja, E., Ossipov, V., Ossipova, S. and Pihlaja, K. (2000). Covariation of fluctuating asymmetry, herbivory and chemistry during birch leaf expansion. *Oecologia* **122**, 354–360.

Leung, B. and Forbes, M.R. (1996). Fluctuating asymmetry in relation to stress and fitness: effects of trait type as revealed by meta-analysis. *Ecoscience* **3**, 400–413.

Leung, B., Forbes, M.R. and Houle, D. (2000). Fluctuating asymmetry as a bioindicator of stress: comparing efficacy of analyses involving multiple traits. *Am. Nat.* **155**, 101–115.

Liermann, M. and Hilborn, R. (1997). Depensation in fish stocks: a hierarchic Bayesian meta-analysis. *Can. J. Fisheries Aqua. Sci.* **54**, 1976–1984.

Light, R.J. and Pillemer, D.B. (1984). *Summing Up: The Science of Reviewing Research.* Harvard University Press, Cambridge, MA.

Mann, C. (1990). Meta-analysis in the breech. *Science* **249**, 476–480.

Mann, C. (1994). Can meta-analysis make policy? *Science* **266**, 960–962.

Marchant, L.F. and McGrew, W.C. (1991). Laterality of function in apes – a meta-analysis of methods. *J. Hum. Evol.* **21**, 425–438.

Markow, T.A. and Clarke, G.M. (1997). Meta-analysis of the heritability of developmental stability: a giant step backward – comment. *J. Evol. Biol.* **10**, 31–37.

McCann, K.S., Hastings, A. and Strong, D.R. (1998). Trophic cascades and trophic trickles in pelagic food webs. *Proc. Roy. Soc. Lond. Ser. B.* **265**, 205–209.

McCarthy, M.A. (1997). Competition and dispersal from multiple nests. *Ecology* **78**, 873–883.

McGrew, W.C. and Marchant, L.F. (1997). On the other hand: current issues in and meta-analysis of the behavioral laterality of hand function in nonhuman primates. *Yearbk Phys. Anthropol.* **40**, 201–232.

Medlyn, B.E., Badeck, F.-W., de Pury, D.G.G., Barton, C.V.M., Broadmeadow, M., Ceulemans, R., de Angelis, P., Forstreuter, M., Jach, M.E., Kellomäki, S., Laitat, E., Marek, M., Philippot, S., Rey, A., Strassemeyer, J., Laitinen, K., Liozon, R., Portier, B., Roberntz, P., Wang, K. and Jarvis, P.G. (1999). Effects of elevated [CO_2] on photosynthesis in European forest species: a meta-analysis of model parameters. *Plant Cell Envir.* **22**, 1475–1495.

Miller, C.K. and Fair, J.M. (1997). Effects of blow fly (*Protocalliphora spatulata*: Diptera: Calliphoridae) parasitism on the growth of nestling savannah sparrows in Alaska. *Can. J. Zool.* **75**, 641–644.

Møller, A.P. (1998). Developmental instability as a general measure of stress. *Stress Behav.* **27**, 181–213.

Møller, A.P. (1999). Developmental stability is related to fitness. *Am. Nat.* **153**, 556–560.

Møller, A.P. and Alatalo, R.V. (1999). Good-genes effects in sexual selection. *Proc. Roy. Soc. Lond. Ser. B.* **266**, 85–91.

Møller, A.P. and Ninni, P. (1998). Sperm competition and sexual selection: a meta-analysis of paternity studies of birds. *Behav. Ecol. Sociobiol.* **43**, 345–358.

Møller, A.P. and Shykoff, J.A. (1999). Morphological developmental stability in plants: Patterns and causes. *Int. J. Plant Sci.* **160**, S135–S146.

Møller, A.P. and Thornhill, R. (1997a). A meta-analysis of the heritability of developmental stability. *J. Evol. Biol.* **10**, 1–16.

Møller, A.P. and Thornhill, R. (1997b). Developmental instability is heritable – reply. *J. Evol. Biol.* **10**, 69–76.

Møller, A.P. and Thornhill, R. (1998). Bilateral symmetry and sexual selection: a meta-analysis. *Am. Nat.* **151**, 174–192.

Mosquera, I., Cote, I.M., Jennings, S. and Reynolds J.D. (2000). Conservation benefits of marine reserves for fish populations. *Anim. Cons.* **3**, 321–332.

Murray, B.R. (1998). Density-dependent germination and the role of seed leachate. *Aust. J. Ecol.* **23**, 411–418.

Myers, R.A. (1998). When do environment-recruitment correlations work? *Rev. Fish Biol. Fisheries* **8**, 285–305.

Myers, R.A. and Mertz, G. (1998). Reducing uncertainty in the biological basis of fisheries management by meta-analysis of data from many populations: a synthesis. *Fisheries Res.* **37**, 51–60.

Myers, R.A., Bradford, M.J., Bridson, J.M. and Mertz, G. (1997). Estimating delayed density-dependent mortality in sockeye salmon (*Oncorhynchus nerka*): a meta-analytic approach. *Can. J. Fisheries Aqua. Sci.* **54**, 2449–2462.

Osenberg, C.W. and St Mary, C.M. (1998). Meta-analysis: synthesis or statistical subjugation? *Integrative Biol.: Issues, News Views* **1**, 43–48.

Osenberg, C.W., Sarnelle, O. and Cooper, S.D. (1997). Effect size in ecological experiments: the application of biological models in meta-analysis. *Am. Nat.* **150**, 798–812.

Osenberg, C.W., Sarnelle, O. and Goldberg, D.E. (1999a). Meta-analysis in ecology: concepts, statistics, and applications. *Ecology* **80**, 1103–1104.

Osenberg, C.W., Sarnelle, O., Cooper, S.D. and Holt, R.D. (1999b). Resolving ecological questions through meta-analysis: goals, metrics, and models. *Ecology* **80**, 1105–1117.

Palmer, A.R. (1999). Detecting publication bias in meta-analyses: a case study of fluctuating asymmetry and sexual selection. *Am. Nat.* **154**, 220–233.

Palmer, A.R. and Strobeck, C. (1997). Fluctuating asymmetry and developmental stability: heritability of observable variation vs. heritability of inferred cause. *J. Evol. Biol.* **10**, 39-49.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *Br. Med. J.* **3**, 1243–1246.

Pearson, K. (1933). On a method of determining whether a sample of size *n* supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika* **25**, 379–410.

Peipert, J.G. and Bracken, M.B. (1997). Systematic reviews of medical evidence: The use of meta-analysis in obstetrics and gynecology. *Obst. Gynecol.* **89**, 628–633.

Peterman, R.M. (1995). Statistical power of methods of meta-analysis. *TREE* **10**, 460.

Petersen, J.E., Cornwell, J.C. and Kemp, W.M. (1999). Implicit scaling in the design of experimental aquatic ecosystems. *Oikos* **85**, 3–18.

Peterson, A.G., Ball, J.T., Luo, Y., Field, C.B., Reich, P.B., Curtis, P.S., Griffin, K.L., Gunderson, C.A., Norby, R.J., Tissue, D.T., Forstreuter, M., Rey, A., Vogel, C.S. and CMEAL Participants (1999). The photosynthesis–leaf nitrogen relationship at ambient and elevated atmospheric carbon dioxide: a meta-analysis. *Global Change Biol.* **5**, 331–346.

Piegorsch, W.W., Smith, E.P., Edwards, D. and Smith, R.L. (1998). Statistical advances in environmental science. *Stat. Sci.* **13**, 186–208.

Pither, J. and Taylor, P.D. (2000). Directional and fluctuating asymmetry in the black-winged damselfly *Calopteryx maculata* (Beauvois) (Odonata: Calopterygidae). *Can. J. Zool.* **78**, 1740–1748.

Planque, B. and Fredou, T. (1999). Temperature and the recruitment of Atlantic cod (*Gadus morhua*). *Can. J. Fisheries Aqua. Sci.* **56**, 2069–2077.

Pomiankowski, A. (1997). Genetic variation in fluctuating asymmetry – comment. *J. Evol. Biol.* **10**, 51–55.

Poulin R. (1994). Meta-analysis of parasite-induced behavioral changes. *Anim. Behav.* **48**, 137–146.

Poulin, R. (1996). Helminth growth in vertebrate hosts: does host sex matter? *Int. J. Parasitol.* **26**, 1311–1315.

Poulin, R. (2000). Manipulation of host behaviour by parasites: a weakening paradigm? *Proc. Roy. Soc. Lond. Ser. B.* **267**, 787–792.

Proulx, M. and Mazumder, A. (1998). Reversal of grazing impact on plant species richness in nutrient-poor vs. nutrient-rich ecosystems. *Ecology* **79**, 2581–2592.

246J. GUREVITCH *ET AL.*

Raudenbush, S.W. (1994). Random effects models. In: *The Handbook of Research Synthesis* (Ed. by H. Cooper and L.V. Hedges), pp. 301–321. Russell Sage Foundation, New York.

Riessen, H.P. (1999). Predator-induced life history shifts in *Daphnia*: a synthesis of studies using meta-analysis. *Can. J. Fisheries Aqua. Sci.* **56**, 2487–2494.

Rosenberg, M.S., Adams, D.C. and Gurevitch, J. (1997). *Meta Win: Statistical Software for Meta-Analysis with Resampling Tests.* Ver. 1.0. Sinauer Associates, Sunderland, MA.

Rosenberg, M.S., Adams, D.C. and Gurevitch, J. (2000). *Meta Win: Statistical Software for Meta-Analysis with Resampling Tests.* Ver. 2.0. Sinauer Associates, Sunderland, MA.

Rosenthal, R. (1979). The 'file drawer problem' and tolerance for null results. *Psychol. Bull.* **86**, 638–641.

Sacks, H.S., Berrier, J., Reitman, D., Ancona-Berk, V.A. and Chalmers, T.C. (1987). Meta-analyses of randomized controlled trials. *New Engl. J. Med.* **316**, 450–455.

Schalk, G. and Forbes, M.R. (1997). Male biases in parasitism of mammals: effects of study type, host age, and parasite taxon. *Oikos* **78**, 67–74.

Schmitz, O.J., Hamback, P.A. and Beckerman, A.P. (2000). Trophic cascades in terrestrial systems: A review of the effects of carnivore removals on plants. *Am. Nat.* **155**, 141–153.

Sharp, D. (1997). Of apples and oranges, file drawers and garbage: why validity issues in meta-analysis will not go away. *Clin. Psych. Rev.* **17**, 881–901.

Silvertown, J. and Dodd, M. (1996). Comparing plants and connecting traits. *Phil. Trans. Roy. Soc. Lond.* **351**, 1233–1239.

Simmons, L.W., Tomkins, J.L., Kotiaho, J.S. and Hunt, J. (1999). Fluctuating paradigm. *Proc. Roy. Soc. Lond. Ser. B.* **266**, 593–595.

Simpson, G.G., Roe, A. and Lewontin, R.C. (1960). *Quantitative Zoology,* revised edn. Harcourt, Brace and Co., New York.

Smith, M.L., Glass, G.V. and Miller, T.I. (1980). *The Benefits of Psychotherapy.* Johns Hopkins Press, Baltimore, MD.

Sokolovska, N., Rowe, L. and Johansson, F. (2000). Fitness and body size in mature odonates. *Ecol. Entomol.* **25**, 239–248.

Strain, B.R. and Cure, J.D. (1994). *Direct effects of atmospheric CO_2 enrichment on plants and ecosystems: an updated bibliographic data base.* Oak Ridge National Laboratory/CDIAC-70, Oak Ridge, TN.

Swaddle, J.P. (1997). On the heritability of developmental stability – comment. *J. Evol. Biol.* **10**, 57–61.

Swales, J.D. (1999). Evidence-based medicine and hypertension. *J. Hypertension* **17**, 1511–1516.

Thornhill, R., Møller, A.P. and Gangestad, S.W. (1999). The biological significance of fluctuating asymmetry and sexual selection: a reply to Palmer. *Am. Nat.* **154**, 234–241.

Thrush, S.F., Schneider, D.C., Legendre, P., Whitlatch, R.B., Dayton, P.K., Hewitt, J.E., Hines, A.H., Cummings, V.J., Lawrie, S.M., Grant, J., Pridmore, R.D., Turner, S.J. and Mcardle, B.H. (1997). Scaling-up from experiments to complex ecological systems: where to next? *J. Exp. Mar. Biol. Ecol.* **216**, 243–254.

Tippett, L.H.C. (1931). *The Methods of Statistics.* Williams & Norgate, London.

Tonhasca, A. and Byrne, D.N. (1994). The effects of crop diversification on herbivorous insects – a meta-analysis approach. *Ecol. Entomol.* **19**, 239–244.

Turchin, P. (1995). Chaos in microtine populations. *Proc. Roy. Soc. Lond. Ser. B.* **262**, 357–361.

Van Zandt, P.A. and Mopper, S. (1998). A meta-analysis of adaptive deme formation in phytophagous insect populations. *Am. Nat.* **152**, 595–604.

VanderWerf, E. (1992). Lack's clutch size hypothesis – an examination of the evidence using meta-analysis. *Ecology* **73**, 1699–1705.

Van Dongen, S., Sprengers, E., Lofstedt, C. and Matthysen, E. (1999). Heritability of tibia fluctuating asymmetry and developmental instability in the winter moth (*Operophtera brumata* L.) (Lepidoptera, Geometridae). *Heredity* **82**, 535–542.

Vasquez, R.A., Bustamante, R.O. and Simonetti, J.A. (1995). Granivory in the Chilean matorral: extending the information on arid zones of South America. *Ecography* **18**, 403–409.

Venier, L.A. and Fahrig, L. (1998). Intra-specific abundance–distribution relationships. *Oikos* **82**, 483–490.

Vollestad, L.A., Hindar, K. and Møller, A.P. (1999). A meta-analysis of fluctuating asymmetry in relation to heterozygosity. *Heredity* **83**, 206–218.

Wand, S.J.E., Midgley, G.F., Jones, M.H. and Curtis, P.S. (1999). Responses of wild C_4 and C_3 grass (Poaceae) species to elevated atmospheric CO_2 concentration: a test of current theories and perceptions. *Global Change Biol.* **5**, 723–741.

Wang, M.C. and Bushman, B.J. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychol. Meth.* **3**, 46–54.

Westoby, M. (1998). A Leaf–Height–Seed (LHS) plant ecology strategy scheme. *Plant Soil* **199**, 213–227.

Whitlock, M.C. and Fowler, K. (1997). The instability of studies of instability – comment. *J. Evol. Biol.* **10**, 63–67.

Windig, J.J. and Nylin, S. (2000). How to compare fluctuating asymmetry of different traits. *J. Evol. Biol.* **13**, 29–37.

Wooster, D. (1994). Predator impacts on stream benthic prey. *Oecologia* **99**, 7–15.

Wortman, P.M. (1994). Judging research quality. In: *The Handbook of Research Synthesis* (Ed. by H. Cooper and L.V. Hedges), pp. 97–110. Russell Sage Foundation, New York.

Xiong, S.J. and Nilsson, C. (1999). The effects of plant litter on vegetation: a meta-analysis. *J. Ecol.* **87**, 984–994.