Ecology, 80(4), 1999, pp. 1142–1149
© 1999 by the Ecological Society of America

STATISTICAL ISSUES IN ECOLOGICAL META-ANALYSES

JESSICA GUREVITCH[1] AND LARRY V. HEDGES[2]

[1] *Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11794-5245 USA*
[2] *Departments of Education, Psychology, and Sociology, The University of Chicago, 5835 South Kimbark Avenue, Chicago, Illinois 60637 USA*

Abstract. Meta-analysis is the use of statistical methods to summarize research findings across studies. Special statistical methods are usually needed for meta-analysis, both because effect-size indexes are typically highly heteroscedastic and because it is desirable to be able to distinguish between-study variance from within-study sampling-error variance. We outline a number of considerations related to choosing methods for the meta-analysis of ecological data, including the choice of parametric vs. resampling methods, reasons for conducting weighted analyses where possible, and comparisons fixed vs. mixed models in categorical and regression-type analyses.

Key words: data synthesis; ecological data, meta-analysis; effect size; heteroscedasticity; meta-analysis; mixed-model analysis; randomization tests in meta-analysis; resampling tests; statistical techniques for data synthesis.

INTRODUCTION

Ecologists concerned with reaching generalizations from a collection of papers reporting experimental results have employed a number of approaches. In recent years, as a large body of experimental data has accumulated to address a number of pressing issues in both basic and applied ecology, ecologists have begun to explore the field of quantitative data synthesis, or meta-analysis. Techniques for meta-analysis were developed in other disciplines, notably the medical, physical, and behavioral sciences (see e.g., National Research Council 1992, Cooper and Hedges 1994), and many of the issues and problems that ecologists are now beginning to grapple with have been confronted in those fields as well. Issues that have been discussed at great length in other fields in which meta-analysis is widely used include approaches to searching the literature, methods for dealing with studies of mixed quality, and publication bias (see *Limitations and problems . . . : Publication bias,* below). Other issues are unique to ecological data. Progress in the quantitative synthesis of ecological data from independent experiments will depend upon appropriate application of the work in meta-analysis that has been done in other research fields, and upon developing new approaches where they are needed.

Meta-analysis offers a subtly different perspective on the outcome of experiments than the one with which ecologists are familiar. Instead of providing a definitive demonstration of a particular phenomenon (a "textbook example"), the outcomes of research studies, like the primary data from which they are derived, are treated as if they are subject to sampling uncertainties. That is, the magnitude of the treatment effect observed in a particular experiment is subject to chance variation. Meta-analysis statistical methods can aid in the summary and interpretation of the findings from a collection of experiments, taking the chance character of those findings into account. Typically, in a meta-analysis the outcome of each study is summarized as an index of effect size and these indices are summarized across studies. Statistical analyses of effect sizes can be constructed to answer a great many questions. For example, how large is the effect overall? Is it positive or negative, and is it reliably different than zero? Are the results consistent across studies? If the results are not in agreement among studies, are there differences in the magnitude of the effect among meaningful categories of studies (e.g., does the effect differ among systems, trophic levels, etc.)?

Once the data are collected, the process of carrying out a meta-analysis typically involves choosing an appropriate metric of effect size (Osenberg et al. 1999), calculating grand-mean effect size across studies and means for different categories of explanatory variables (or slopes, where the explanatory variable is continuous), determining the confidence intervals around the means or slopes, and then carrying out statistical tests to determine the consistency of the effects within and among categories of studies. Each of these steps requires decisions that have both scientific and statistical implications. In this paper, we focus on the statistical issues involved in these decisions.

The choice of an effect-size metric underlies the entire meta-analysis and is therefore a decision of critical importance (Osenberg et al. 1999). The validity of statistical tests and confidence intervals in meta-analysis

Manuscript received 11 November 1997; revised 2 July 1998; accepted 8 July 1998; final version received 11 September 199. For reprints of this Special Feature, see footnote 1, p. 1103.

depends on accurate calculation of the variances (or standard errors) of effect estimates. This depends upon understanding the statistical properties (the sampling distribution) of the effect-size metric one chooses, and the ability to calculate the standard error of the effect-size estimate. For standard metrics of effect size, such as the standardized mean difference (d), the z transform of the correlation coefficient, and the log odds ratio, the sampling variances are well known. For metrics of effect size whose properties are not well known, the distributions must be derived to obtain these variances (see e.g., Hedges et al. 1999). Other measures of effect size may be biologically meaningful, but the sampling variances may be very difficult to derive, potentially restricting their usefulness in meta-analyses. The calculation of sampling variances usually depends upon having information about sample sizes and some sort of variance estimate for each study. However, for some metrics (such as d) a reasonable approximation of the sampling variance can be calculated using sample sizes alone (see e.g., Hedges and Olkin 1985).

Ecologists have been reviewing and synthesizing published results without any special statistical techniques for a long time. Is the introduction of unfamiliar techniques really needed? For example, one might evaluate the importance of an effect by assembling the publications that test a hypothesis, counting the number of statistically significant outcomes, and deciding that the effect exists or is important if the proportion of studies that found the effect is substantial compared with the proportion that did not detect a significant result. Statistical tests might be used to compare the number of studies with significant and nonsignificant outcomes. This approach, called "vote counting," has historically been widely used to summarize results from multiple studies and is still fairly common in ecological publications (e.g., Hartley and Hunter 1998). Unfortunately, vote counting has very poor properties as a statistical procedure. The results of vote counts are seriously biased, the method has low statistical power, and most importantly it fails to provide critical information on the overall results of the body of studies (Hedges and Olkin 1985). It is interesting to note that not only are the problems of vote-count procedures not ameliorated by including a greater number of studies in one's synthesis, but that the power of this technique actually decreases as more information (more studies) are available, tending to zero as the number of studies becomes large (Hedges and Olkin 1980, 1985).

Similarly, it is also common to decide that the results of different experiments are consistent with one another (e.g., "Smith replicated the findings of Jones") if both studies obtain the same outcome of a significance test for the effect of the treatment, and inconsistent ("Smith failed to replicate") if the outcomes of significance tests differ. Here, too, the intuition of many researchers is incorrect. The chance that the outcome of two experiments with identical underlying effects agree on the outcome of significance tests for those effects is $p^2 + (1 - p)^2$, where p is the statistical power of the test in each study (here assumed to be equal). Given a reasonably powerful statistical test ($p = 0.8$), this chance of agreement is, non-intuitively, only 68% when both experiments are examining identical underlying effects.

Properties of Meta-analysis Data

The statistical methods that are useful for meta-analysis are laid out in considerable detail elsewhere (e.g., Hedges and Olkin 1985, *in press*, Gurevitch and Hedges 1993, Cooper and Hedges 1994) and we will not repeat those details here. Rather, we focus on how to evaluate different approaches in undertaking a meta-analysis of ecological data. Researchers might, for example, be aware of the problems associated with vote counting, but still consider analyzing effect sizes as if they were measurements made on individuals, carrying out regression, ANOVA, etc., on the effect-size estimates derived from the various studies that they wished to synthesize (e.g., Hartley and Hunter 1998). This approach, however, has serious disadvantages. To understand these disadvantages, we need to examine two features that characterize meta-analysis data sets and that have important implications for statistical analysis.

Two sources of variation

The first feature that is characteristic of even the simplest meta-analysis data set is that the data have a hierarchical structure with at least two sources of variation: within-study sampling error and between-study variation in effect-size parameters. That is, there is a study-specific sampling error (uncertainty) associated with every effect-size estimate in a meta-analysis. This sampling error is the deviation between the effect-size estimate obtained and the study-specific effect-size parameter, the value that would have been obtained if the within-study sample size had been so large that there was essentially no sampling error. There may also be (and typically are) real differences between studies in the magnitude of the effect sizes, which gives rise to a second component of variation due to between-study differences. Thus the observed effect-size estimates can be conceived as having two components of variation, one associated with study-specific sampling errors and another associated with between-study differences in the underlying study-specific effect-size parameters.

The first source of variation may be of little scientific interest, but is important to account for statistically. One can think of this term as quantifying the variation that would be obtained if a given experiment were replicated, in exactly the same way, with a different sample of replicates (e.g., different individual plots or organisms). The second source of variation, variation of true effects, may be of considerable scientific interest, particularly if it is associated with important characteristics of the experiments (such as species, treatment intensity, or experimental procedure).

Unequal error variances among studies

A second, related characteristic of meta-analysis data sets is that the study-specific sampling-error variances are almost never identical across studies. The sampling variances of each study are inversely proportional to within-study sample sizes, and sample sizes may vary greatly across studies. For instance, in the data reported by Curtis and Wang (1998), the largest variance was over 1000 times the magnitude of the smallest. Thus, while in a primary study one is concerned with violating statistical assumptions of homogeneity due to variability in the effect, in a meta-analysis, one is particularly concerned with the heterogeneity in the precision of the effect-size estimates, which depends upon heterogeneity of the within-study sample sizes. The resulting heteroscedasticity under many conditions will seriously violate the assumptions of conventional parametric statistical tests such as ANOVA and regression (but see *Limitations and problems . . . : Incomplete reporting . . . ,* below). While primary data in many ecological data sets may also violate this assumption, in primary data variance is often directly related to the size of the effect measured, and in those cases can thus be transformed to meet the homogeneity-of-variances assumption. In meta-analysis the within-study variance is typically not monotonically related to the size of the effect, and thus cannot be eliminated by data transformation. More fundamentally, the usual approaches to data transformation would only resolve the second source of heterogeneity of variances described above (due to variation in effect sizes among studies) and would not generally account for the first and statistically far more problematic source of variance heterogeneity (that due primarily to differences in sample sizes among studies).

Neither of the features discussed above is unique to meta-analysis data, and in fact the data from some primary studies may share both of them. For example, multisite studies with different numbers of replicates in each site may exhibit both between- and within-site variation and also substantially different within-site variation. In such multisite studies the problems of primary statistical analysis closely resemble those of meta-analysis.

Weighting to Account for Unequal Variances

One of the most useful solutions when faced with the heteroscedasticity typically found in meta-analysis data is to weight effect sizes for statistical analysis by the inverse of the sampling variance of the effect size. Most statistical methods for meta-analysis employ weighting, which really has two purposes. First, weighting when computing means (or regression coefficients, etc.) increases the precision of the combined estimates and increases the power of tests. These increases in precision and power are not negligible: increases in power of 50–100% in weighted, vs. unweighted, tests of the significance of the mean can easily happen. Note that this difference in power does not compromise robustness when the null hypothesis is true.

A second purpose of weighting is that it makes certain statistics have simpler sampling distributions. This is merely a technical advantage of weighting, but an important practical one since it means that test statistics have standard sampling distributions (like chi-squares) whose critical values have been tabulated and thus no special tables or numerical algorithms are necessary. Moreover, (at least in large samples), the weighted combined estimates and tests have optimal properties. The tests are the most powerful in a wide class of tests of the relevant hypotheses, and the estimates are more precise than any others that could be computed (Hedges 1983*b*, Hedges and Olkin, 1985).

A common misconception is that weighting has something to do with bias. Weighting is *not* necessary to reduce bias. If the individual estimates are unbiased, the unweighted mean is unbiased too. However weighting does have the desirable property (in addition to the increase in precision) of counting large studies more heavily than small ones, which often seems reasonable in summarizing overall results.

Statistical Models for Categorical Meta-analysis

What methods are most appropriate, powerful, and informative for the analysis of weighted meta-analysis data? Statistical methods for combining estimates across experiments have a long history, dating at least from work by Cochran (1937). Such methods generally fall into one of three categories: fixed-, mixed-, or random-effects procedures. In fixed-effect models, it is assumed that all studies with similar-enough characteristics share a common, "true" effect size, and estimates differ from one another by sampling error only (see e.g., Fleiss 1981, Hedges 1982*a*, Rosenthal and Rubin 1982: chapter 10). In a random-effects model, the true effect size is expected to differ among studies, and the goal of the analysis is to quantify the variation in the effect parameters (see e.g., Hedges 1983*a*, DerSimonian and Laird 1986). Random-effects models are intuitively more appealing in ecology than fixed-effects models because we would often expect the true effect to vary among studies. The typical form of random-effects analyses is described in the context of response ratios in Hedges et al. (1999).

One important limitation of pure random-effects models is that they do not provide a way to determine how effects may depend upon important substantive characteristics of studies. Analyses using fixed-effect models can do so in a straightforward way by accounting for variation among categories of studies, which is often what will be most interesting in an ecological meta-analysis (e.g., Hedges 1982*b*). However, the scientifically interesting categories of studies may still contain substantial real variation in effect size, which makes fixed-effects models inappropriate.

Gurevitch and Hedges (1993) and Stram (1996) have proposed the use of mixed models in meta-analysis as a way to combine the advantages of random- and fixed-effects models, much as in the analysis of primary data. Mixed models are appropriate for analyzing differences between groups of experiments when the groups are not expected to be internally homogeneous. The distinction between fixed, random, and mixed models are familiar to many ecologists in conventional analysis of variance. Although the details of mixed models in meta-analysis differ from those in ANOVA, the underlying principles are similar. In essence the effects of the groupings of experiments on effect parameters are fixed effects while the variations among the effect parameters of experiments within groupings are taken to be random effects. The details of both simple and more complex approaches to mixed-model meta-analysis are provided elsewhere (Hedges and Olkin, *in press*). One of these approaches was outlined by Gurevitch and Hedges (1993), and has been incorporated in a software package for meta-analysis (Rosenberg et al. 1997). The mixed model is more general than either fixed- or random-effects models, because it collapses to a fixed-effects model if there is no variation left after accounting for differences among categories and for sampling error, and to a random-effects model if the studies all belong to a single category.

An alternative to the parametric tests above are resampling techniques to test for differences among categories of studies (Adams et al. 1997). Resampling can also be used to bootstrap confidence intervals around weighted-mean effect sizes. If the researcher has reason to believe that the data might deviate substantially from the normality and large-sample assumptions of parametric meta-analysis models (e.g., see Hedges and Olkin 1985), randomization tests can be advantageous. In particular, the parametric tests outlined above assume that the data used to calculate effect sizes in each of the component studies (i.e., the data used to calculate the means of the experimental and control groups) are normally distributed. Resampling methods do not assume normality, provide valid sampling distributions for test statistics, and, unlike conventional nonparametric tests such as, e.g., rank tests, randomization tests are not necessarily less powerful than the corresponding parametric tests (Manly 1991:32; assuming that both are carried out on weighted data). However, a potential disadvantage of using randomization tests is that it is not possible to separate the two sources of variance discussed above (within-study sampling error and between-study variation in true effects; see *Properties of meta-analysis data: Two sources of variation*).

Resampling tests can be set up in many different ways. One approach, for example, to testing for differences among classes of studies (high arctic vs. low arctic tundra, carnivores vs. herbivores) is to randomly reassign outcomes to classes (without replacement), then repeat this many times, each time computing the test statistic (here, the homogeneity statistic Q_B; e.g., Hedges and Olkin 1985) to determine a distribution against which to evaluate the significance level of the actual value of the test statistic. Since differences among the effect-size estimates of studies (and not just their sampling errors) contribute to variation in the test statistic, resampling tests are analogous to mixed models, which also incorporate these two sources of variation into the uncertainty of test statistics.

Regression-type Approaches

Not all ecological meta-analysis fits an ANOVA-like framework. In many meta-analyses, a regression-type approach is demanded. While regressions have been carried out in meta-analysis in other disciplines, this area is still a new one in ecology. As in any meta-analysis, simply taking meta-analytic data and using conventional (unweighted) regression techniques has drawbacks that are outlined above (see *Properties of meta-analysis data: Unequal error variances among studies*). Parametric methods for carrying out regressions on effect sizes are available in the meta-analysis literature (e.g., see Hedges and Olkin 1985:167–188, Raudenbush and Bryk 1985, Cooper and Hedges 1994: 295–299, 308–321).

These methods involve weighting and require that estimates of within-study sampling-error variances are available. The fixed-effects regression approach involves weighting each effect size by the reciprocal of its sampling-error variance. The regression coefficients (i.e., slopes) given by standard weighted-regression programs are correct for this analysis but the standard errors of the slopes must be "corrected" by dividing them by the square root of the mean squared error for the analysis of variance of the regression (the square root of the error variance). The test statistic for testing that the regression coefficient is different from zero is the ratio of the regression coefficient divided by this corrected standard error, and is compared to the critical values of the standard normal distribution (Hedges and Olkin 1985).

A mixed-model regression approach can be implemented in several ways (Hedges 1992*a*). One simple way to do this requires several steps. First compute a preliminary unweighted regression to obtain an estimate of the residual variance component. The residual variance component estimate is the difference between the mean squared error for the analysis of variance of the regression and the average of the sampling error variances from each of the studies. Then compute the mixed-model weights by taking the reciprocals of the sum of the sampling-error variance and the residual variance component. Finally carry out a weighted analysis in exactly the same way as that of the fixed-effects analysis, except using the mixed-model weights. While this will provide a workable analysis, more sophisticated analyses (e.g., involving maximum-likelihood estimation) provide more efficient estimation and more powerful tests of hypotheses about regression coefficients (e.g., Hedges and Olkin, *in press*).

Limitations and Problems with Ecological Data in Meta-analysis

While meta-analysis offers powerful tools to address many important questions in ecology, like any other statistical approach, it is subject to various limitations. It is also subject to a number of as-yet-unresolved problems. Four issues that are of particular concern to ecologists are: incomplete data reporting, the lack of independence among effect-size estimates, publication bias, and research bias. We believe strongly that by far the most serious of these problems, and the one that is in principle easiest for ecologists to do something about, is poor data reporting.

Incomplete reporting of primary studies and missing data in meta-analysis

A serious problem for the wide application of meta-analytic methods in ecology is poor reporting of ecological data (e.g., see discussion in Gurevitch et al. [1992]). Ecological experiments commonly fail to report sample sizes and variances, for instance, which makes it impossible to include those studies in a meta-analysis that uses the standard (weighted) parametric statistical tests designed for meta-analysis. The obvious solution is to upgrade publication standards, alerting authors, reviewers, and editors so that papers are not published without the basic information necessary for readers to properly evaluate the results (see e.g., *Ecology* **79**(1):"Instructions to Authors"; see "Guidelines for Statistical Analysis and Data Presentation" on the Ecological Society of America home page). Lofty aspirations aside, are there any alternatives, particularly for literature that has already been published? While it is preferable to do conventional, weighted meta-analysis, including only those studies with satisfactory data reporting, can anything be done if it is impossible to weight the effect sizes because most studies omit variance and sample-size information needed to compute the standard error of effect-size estimates (e.g., Downing et al. 1999)?

We argue that, given a body of literature, some information regarding the overall findings is much better than no information, and that therefore it is desirable to develop methods for data synthesis of poorly reported data (e.g., where no estimate of sampling variance is published). There are two approaches that one can take to carry out the meta-analysis: unweighted standard parametric statistical tests (such as ordinary least-squares regression or ANOVA) or unweighted randomization tests. Neither is ideal, but where there is no alternative, they may provide useful information where otherwise none is available.

Both randomization tests and unweighted parametric statistical methods assume homogeneity of variances (HOV). However, violation of the HOV assumptions of both conventional statistical methods (such as unweighted ANOVA and least-squares regression) and unweighted randomization tests do not always seriously compromise the Type I error rates of these tests. The exact consequences of the violation of the HOV assumption depend upon the nature of the data structure for both unweighted randomization and parametric approaches (see e.g., Cressie and Whitford 1986, Stewart-Oaten 1995). When the error variances are not confounded with characteristics of substantive interest in the analysis, then either conventional statistical methods or resampling methods may provide fairly reliable (if suboptimal) results. For example, when testing for differences among groups of studies using ANOVA, if the pattern (i.e., the distribution) of sampling-error variances does not differ substantially among groups, or if the pattern of sampling-error variances is the same for each value of the independent variable in a regression analysis, then the Type I error rate should be close to the intended value. On the other hand, when, for example, the patterns of error variances differ among groups of studies in an ANOVA, or increase with increasing values of the independent variable in a regression, the Type I error rate obtained from the analysis may be highly misleading. (For analytic results on the effects of unequal variances in analyses based on the general linear model, see Goldgerber [1964:238–241].) In all cases, the power and the Type II error rate will be affected and parameters will be estimated with less precision, with the weighted analysis always being more powerful.

Adams et al. (1997) and Rosenberg et al. (1997) recently suggested one approach to using resampling tests on unweighted data in meta-analysis. This approach allows one to derive confidence limits on mean effect-size estimates and evaluate the statistical significance of differences between the effect sizes of groups of studies, assuming that the primary research papers publish sufficient data to obtain effect-size estimates. Confidence intervals can be obtained by bootstrapping the (unweighted) effect-size data, while homogeneity tests can be constructed as outlined above for weighted data. As above, this approach could yield either reliable or invalid results, depending upon the nature of the error structure of the data.

Similarly, resampling methods can also be used to obtain tests associated with regression of effect-size estimates on some continuous predictor variables. The most straightforward approach in this case is to calculate the slope using ordinary unweighted least-squares regression. The statistical significance of the slope can then be determined by randomly reassigning the dependent and independent variables (without replacement), respectively, in each pair (the effect sizes and the continuous variable) many times, each time recording the value of the slope. The actual slope is then judged against the distribution of slopes thus obtained. A confidence interval around the slope can be determined by bootstrapping, if needed. Each of these resampling approaches should be viewed as analogous to mixed-model analyses, because they take heterogeneity of effects into account in determining uncertainty of mean effects, but use equal weights.

It is not clear whether resampling methods are any more robust than conventional unweighted parametric statistical methods to violation of the homogeneity assumption, but they will at least have the advantage of being free of the normality assumptions of the parametric statistical tests. The potential uses of randomization approaches bear further exploration for ecologists interested in quantitative data synthesis.

Non-independence

Non-independence of effect-size estimates is also a problem that frequently arises in meta-analysis. The best ways to deal with it (as in analyses of primary data) remain open to debate and exploration. There are two fundamentally different kinds of non-independence, one associated with each source of variation in meta-analytic data. Both types of non-independence can lead to underestimates of the standard error of the mean effect and therefore liberal evaluations of the statistical significance of effects.

If several different measurements are made on each replicate in a study (e.g., measures at several points in time or of slightly different outcomes) and different effect sizes are computed from each, the different effect sizes may be correlated because the data on which they are based are correlated. This type of dependence arises through the correlations among the within-study sampling errors and can be eliminated by using only one effect-size estimate from each set of replicates, although this approach may involve discarding at least some of the potentially relevant data provided in a study. One alternative is to conduct a different meta-analysis for each kind of effect measure. For example, a set of studies might report on the outcome of competition in terms of both effects on growth and effects on survival, and one could then do a meta-analysis on growth effects and a separate meta-analysis on survival effects. Other (more complex) approaches can sometimes be employed that permit the use of all of the data via multivariate methods that explicitly model the dependence structure (see e.g., Hedges and Olkin 1985).

A different kind of dependence arises in connection with between-study variation of effects. For example, when several effect-size estimates are computed from data from the same laboratory, there may be dependence if common materials, procedures, etc., in a laboratory make the outcomes of separate experiments obtained from the same laboratory less variable than those obtained from different laboratories. Such dependence also arises if there are differences in the effect sizes obtained in studies on different species, but those differences are smaller among closely related species. This might lead to less variability among the effect sizes obtained among studies conducting experiments on similar organisms compared to studies conducting experiments on species that are less closely related (e.g., Harvey and Pagel 1991). One approach to dealing with this kind of data structure is to explicitly model the hierarchical dependence of the effect sizes (e.g., Cheverud et al. 1985, Felsenstein 1985). For example, if experiments are nested within laboratories, then we would estimate components of variance within and between laboratories, and compute the variance of the overall mean effect size as a function of these two variance components.

Publication bias

Publication selection is the tendency for results that are statistically significant to be more likely to be published than those that fail to detect significance (see e.g., Rosenthal 1979, Cooper and Hedges 1994). If severe publication selection exists, it can substantially bias estimates of observed (published) effect sizes (see e.g., Hedges and Olkin 1985: chapter 14). It is important to recognize that this bias is just as much a threat to the validity of interpretation of all published work, including ordinary narrative reviews and even single studies considered in isolation, as it is to meta-analysis. While the extent of publication bias is, given its nature, difficult to study, there is some evidence from the field of medicine that publication bias may not inevitably be as severe as might be imagined (K. Dickersin and Y-I. Min, *unpublished manuscript*). Several approaches to the detection and quantification of publication bias are available, but they are not necessarily all ideal for typical ecological data.

One class of methods for detecting publication bias is based on examination of the relation between standard error and effect size. These include graphical methods (the funnel plot, Light and Pillemer 1984), and formal tests for the correlation between sample size (or standard error of the effect-size estimate) and effect size (Begg and Mazumdar 1994). These methods are sensible if (as in the case of medical experiments evaluating the same treatment on the same types of patients) the effect parameters are expected to be consistent across studies. The rationale of the methods is that, since the effect parameters are consistent across studies, any relation with sample size (or standard error) must reflect publication selection.

However, in many ecological applications, the effects are often expected to vary substantially across experiments. If this is the case, it is good design practice to use larger sample sizes in experiments where effects are expected to be smaller. Consequently, a relation between sample size and effect size may reflect rational experimental design rather than publication bias. In such situations, which may be typical of experimental ecology, tests for publication bias based on more elaborate selection models (Hedges 1992*b*), and procedures for establishing robustness of results (the fail-safe N, Rosenthal 1979) may be more appropriate. In any event one conceivable benefit of discussions about publication bias is that they may prompt greater realization that studies failing to find significant results may provide valuable information that warrants publication.

Research bias

A related issue that appears to be of particular concern to ecologists (J. Gurevitch, *personal observation*) we term "research bias"—the tendency to perform experiments on organisms or under conditions in which one has a reasonable expectation of detecting statistically significant effects. Research bias has not been discussed much in the general meta-analysis literature, and is important with respect to the extent to which one wishes to generalize the results of a body of experimental data. If the results of a meta-analysis are taken to indicate something about the natural world in general, this could potentially result in serious misinterpretation of the conclusions of a meta-analysis. If the results are interpreted to be a summary of the existing experimental data on a problem, the bias may be much less serious.

General Recommendations

We recommend that weighted parametric mixed-model analysis be used as a first choice when sufficient data are reported to compute standard errors of individual effect-size estimates and when the parametric assumptions (e.g., sufficiently large enough within-study sample sizes for estimates to be normally distributed and a normal distribution of between-study effects) are not seriously violated. If the analyst believes that these assumptions of the parametric tests are unlikely to be met, weighted resampling methods are the approach of choice. If insufficient data are reported to calculate standard errors of effect-size estimates, then unweighted resampling methods and methods appropriate for the analysis of primary data (ANOVA, regression, etc.) may sometimes provide valid answers, albeit with reduced power.

Acknowledgments

We appreciate the very thoughtful comments of Allen Stewart-Oaten, Deborah Goldberg, Craig Osenberg, and an anonymous reviewer. They have improved the paper considerably. This work was initiated under a grant from NSF to J.G. (BSR-9113065), and this paper was prepared as part of the Meta-Analysis Working Group supported by the National Center for Ecological Analysis and Synthesis (NCEAS), a Center funded by NSF (Grant no. DEB-94-21535), the University of California–Santa Barbara, and the State of California. We gratefully acknowledge the stimulating atmosphere provided by NCEAS, the Meta-analysis working group, and the NCEAS staff, and particularly thank Mark Schildhauer. This is contribution number 1023 in Ecology and Evolution at the State University of New York at Stony Brook.

Literature Cited

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

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

Cheverud, J. M., M. M. Dow, and W. Leutenegger. 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution **39**:1335–1351.

Cochran, W. G. 1937. Problems arising in the analysis of a series of similar experiments. Journal of the Royal Statistical Society Supplement **4**:102–118.

Cooper, H., and L. V. Hedges. 1994. The handbook of research synthesis. The Russell Sage Foundation, New York, New York, USA.

Cressie, N. A. C., and H. J. Whitford. 1986. How to use the two-sample *t*-test. Biometrical Journal **28**:131–148.

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

Der Simonian, R., and W. Laird. 1986. Meta-analysis in clinical trials. Controlled Clinical Trials **7**:177–188.

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

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist **125**:1–15.

Fleiss, J. 1981. Statistical methods for rates and proportions, Second Edition. John Wiley & Sons, New York, New York, USA.

Goldberger, A. S. 1964. Econometric theory. J. Wiley, New York, New York, USA.

Gurevitch, J., and L. V. Hedges. 1993. Meta-analysis: combining the results of independent studies in experimental ecology. Pages 378–398 *in* S. Scheiner and J. Gurevitch, editors, The design and analysis of ecological experiments. Chapman & Hall, New York, New York, USA.

Gurevitch, J., L. V. Morrow, A. Wallace, and J. A. Walsh. 1992. A meta-analysis of field experiments on competition. American Naturalist **140**:539–572.

Hartley, M. J., and M. L. Hunter, Jr. 1998. A meta-analysis of forest cover, edge effects, and artificial nest predation rates. Conservation Biology **12**:465–469.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford, UK.

Hedges, L. V. 1982*a*. Estimating effect size from a series of independent experiments. Psychological Bulletin **92**:490–499.

———. 1982*b*. Fitting categorical models to effects sizes from a series of experiments. Journal of Educational Statistics **7**:119–137.

———. 1983*a*. A random effects model for effect sizes. Psychological Bulletin **93**:388–395.

———. 1983*b*. Combining independent estimators in research synthesis. British Journal of Mathematical and Statistical Psychology **36**:121–131.

———. 1992*a*. Meta-analysis. Journal of Educational Statistics **17**:279–296.

———. 1992*b*. Modeling publication selection effects in random effects models in meta-analysis. Statistical Science **7**:246–255.

Hedges, L. V., Gurevitch, and P. Curtis. 1999. The meta-analysis using response ratios in experimental ecology. Ecology **80**:1150–1156.

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

Hedges, L. V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, New York, New York, USA.

Hedges, L. V., and I. Olkin. *In press.* Statistical methods for meta-analysis in the medical and social sciences. Academic Press, New York, New York, USA.

Light, R. J., and D. B. Pillemer. 1984. Summing up: the science of reviewing research. Harvard University Press, Cambridge, Massachusetts, USA.

Manly, 1991. Randomization and Monte Carlo methods in biology. Chapman & Hall, London, UK.

National Research Council. 1992. Statistical problems and research opportunities in the combination of information. National Academy of Sciences Press, Washington, D.C., USA.

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

Raudenbush, S. W., and A. S. Bryk. 1985. Empirical Bayes meta-analysis. Journal of Educational Statistics **10**:75–98.

Rosenberg, M. S., D. C. Adams, and J. Gurevitch. 1997. MetaWin: Statistical software for meta-analysis with resampling tests. Version 1.0. Sinauer Associates, Sunderland, Massachusetts, USA.

Rosenthal, R. 1979. The "file drawer problem" and tolerance for null results. Psychological Bulletin **86**:638–461.

Rosenthal, R., and D. B. Rubin. 1982. A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology **74**:166–169.

Stewart-Oaten, A. 1995. Rules and judgements in statistics: three examples. Ecology **76**:2001–2009.

Stram, D. O. 1996. Meta-analysis of published data using a linear mixed-effects model. Biometrics **52**:536–544.