Meta-Analysis: Synthesis or Statistical Subjugation?

Scientists in all disciplines of biology are concerned with summarizing the state of knowledge within their fields. Entire journals (e.g., *Annual Review of Ecology and Systematics, Quarterly Review of Biology*) and national research centers (e.g., The National Center for Ecological Analysis and Synthesis [NCEAS]) have been established based on the assumption that the synthesis of results from separate studies is essential to scientific progress. Clearly, synthesis and review are critical parts of the scientific process, and should offer as much scholarly reward as primary investigations. Unfortunately, despite the importance of synthetic endeavors there has been relatively little development of quantitative methods applicable to synthetic reviews. As a result, reviews have been largely ad hoc narratives, with subjective and qualitative summaries. This lack of formal, rigorous protocol for synthetic research contrasts markedly with the standard quantitative training that biologists receive for individualistic primary research (e.g., training in experimental design and analysis of variance). Fortunately, this asymmetry in tools for primary vs. synthetic investigations has begun to change dramatically in the past 15 years, as meta-analysis has become a common tool in many scientific disciplines.

Meta-analysis is a statistical approach to the analysis of a collection of results from individual studies for the purpose of integrating the findings. Beyond this straightforward definition, meta-analysis means a variety of things to different people. To some, it includes all forms of research synthesis; to others, it is restricted to experimental data; to everyone, it is at least quantitative. Similarly, to some, meta-analysis is a recent invention that defines a revolutionary new research field, whereas to others, it has a long tradition dating back to the early 1900s and is nothing more than another in the vast array of scientific investigatory tools.

Despite the lack of consensus about its specific meaning, there has been a recent barrage of interest in quantitative reviews involving meta-analysis, especially since the publication of Hedges and Olkin's classical book.[1] Nowhere has this impact been more intense than in medicine, where in 1996 alone there were over 600 papers catalogued in Medline that used or discussed meta-analysis. A similar pattern, albeit on a smaller scale, is seen in ecology and evolution, where the first meta-analysis appeared in 1991.[2] In the following 6 years, over two dozen meta-analytic papers were published, a symposium on meta-analysis was sponsored by the Ecological Society of America, and the Meta-Analysis Working Group was sponsored at NCEAS. Some of this interest came, as it often does, from the coining of a new and catchy term—meta-analysis was first used by Glass[3] in 1976—and by the transfer of insights from one discipline to another. On a more substantive level, this interest arose because meta-analysis implied a detailed specification of the statistics and protocol necessary to conduct a rigorous quantitative review. Accompanying this intense interest came a frenzy to jump on the bandwagon and quickly apply this tool to new questions. Indeed, in some contexts, it seemed that a meta-analytic industry was developing to mine the literature and churn out papers using rote techniques. Standard meta-analytic protocols were fast becoming cookbook procedures, despite the cautions of many practitioners and the vast complexities of the technique. The basic recipe is dangerously simple: extract estimates of the magnitude of effect (e_i) and its variance ($v(e_i)$) from each study, and use weighted statistical analyses (e.g., weight by $1/v(e_i)$) to estimate categorical means, or response surfaces, and their confidence limits.

Of course, meta-analysis is no panacea and there are several reasons for expressing caution about its promise. Probably the biggest problem entails lack of independence and bias: e.g., the studies selected for synthesis are inevitably a biased sample of those that have been conducted, and the studies that were conducted are a biased sample of those that could have been conducted. There is no doubt that all environments and taxa are not sampled with equal probability (e.g., there are disproportionately large numbers of studies of endocrine function in white rats and goldfish, species interactions in the rocky intertidal, and trophic cascades in north temperate lakes). There is also little doubt that studies showing significant effects are more likely to get published than studies on "no effect." However, if we are going to condemn meta-analysis on the grounds of publication and study bias, then we have to reject all synthetic reviews—all academic reviews must work with the same set of studies and must therefore be condemned by the same criteria. Given the importance of synthetic reviews, discarding them entirely is hardly a tenable proposition; thus, biologists need better tools that will aid in performing quantitative syntheses.

METAWIN: STATISTICAL SOFTWARE FOR META-ANALYSIS WITH RESAMPLING TESTS

A new software package, *MetaWin*,[4] will greatly accelerate the introduction of meta-analytic procedures to biologists. It is a simple, easy to understand software package that facilitates certain forms of meta-analysis, such as those advocated in the most influential of the meta-analytic papers in ecology and evolution.[5–7] Installation is trivial (it can even be run from a 3.5 in. floppy drive), the windows are simple and well organized, the output is straightforward (with nice features like the path to the source data file and the date and time at which the analysis was run), and the documentation is helpful without going into extraneous detail (i.e., it is brief). Indeed, as testament to the simplicity of the program, the how-to portion of the manual is only 19 pages long. The biggest section of the manual is a reprinting of Gurevitch and Hedges'[6] description of meta-analytic techniques and their application to ecology.

Most meta-analyses proceed in four basic steps: 1) define the question (which includes selecting relevant studies and choosing a metric that summarizes the result of each study); 2) extract and catalogue the data (i.e., quantitative estimates of"effect size" and its variance along with associated covariates or coding variables); 3) analyze the data; and 4) draw biological inferences. *MetaWin* addresses step 3, and largely borrows from statistical protocol developed in the educational and social sciences.[1] It will aid anyone interested in performing meta-analyses; however, like most statistical packages, or any scientific tool for that matter, the results that are generated are only as good as the ideas that led to the analyses. In discussing *MetaWin* and evaluating its usefulness, two important issues are 1) the choice of a metric of effect size, and 2) the related issue of null hypothesis testing.

Metrics of Effect Size

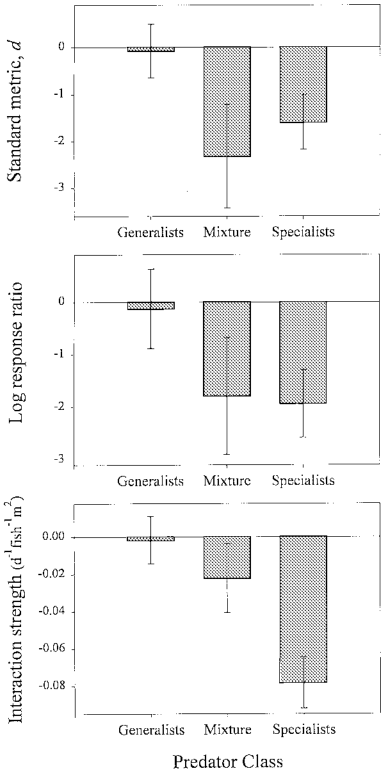
If the results of an experimental or observational study are going to be reduced to some concise quantitative summary index, one must choose a metric of effect size that measures how much a response variable was influenced by an experimental manipulation or naturally varying factor. The classic meta-analytic literature and *MetaWin*, as well as most recent applications in ecology and evolution, rely on a single class of metrics of effect size,[8] the most common of which is the difference in means between two treatment groups divided by the pooled standard deviation. This metric is often referred to as Cohen's *d* or Hedges' *g*, and is one of a family of closely related metrics proposed by Glass.[3] These metrics have been criticized by both statisticians[9] and ecologists.[8,10] Indeed, in his 1995 essay, "A Statistician Looks at Met-Analysis," Finney[9] noted: "I too am surprised that anyone should advocate this *g* as a general measure of effect, independent of context and of practical interpretation… Why then did Glass introduce this *g*, and why do Hedges and Olkin base most of their book upon its use… Hedges and Olkin demonstrate some convenience for *g* on account of its mathematical tractability, but is not intelligible interpretation more important?" Finney concluded that "at present I can see no practical merit in Glass' definition," and conjectures that its use might be "related to the deification of tests of significance as the main objectives of statistical analysis, and consequent neglect of the need to express conclusions on a scale that is biologically meaningful." Thus, Finney highlighted two problems of relevance to the role of meta-analysis in biology: 1) the reliance on biologically ambiguous estimates of effect size[8] and 2) the preoccupation with null hypothesis tests.[11–13] *MetaWin* simultaneously avoids and embraces both criticisms.

One of the best features of *MetaWin* is that it is sufficiently flexible to permit the use of metrics other than *d*. The user needs simply to define the metric and extract an estimate of the effect size and its variance (when possible) from each study. *MetaWin* does not (nor should it, at this time) provide an exhaustive drop-down list of various metrics—it correctly leaves this task to the user. Unfortunately, *MetaWin* does offer two drop-down choices: *d* is the default metric and ln*R* (the log of the ratio of the two treatment means) is the single alternative. These options almost suggest that the user might not need anything else. Indeed, the authors caution against using a user-defined metric without first knowing its statistical properties. We would caution similarly that the user not use *d* or ln*R* without understanding their conceptual link to the questions being explored (e.g., see Box 1).

Indeed, despite the flexibility in choosing an effect size metric, *MetaWin* (as well as much of the meta-analytic field) is deeply entrenched in the *d* culture. Thus, *MetaWin*'s default metric is *d*, and the manual is written almost entirely with reference to *d* and the literature built up around *d*. For example, in the resampling procedures, *MetaWin* allows the user to choose a non-parametric weighting function. The manual is not clear regarding the implications of this choice; however, justification for this function is based on theoretical arguments made for the variance of *d*,[1] and thus is not necessarily appropriate for other metrics (although it is applied to all effect sizes when the non-parametric option is chosen). Unfortunately, the manual does not discuss whether other weighting functions can (or should) be imposed when using metrics other than *d*. As another example of this rather narrow view, chapter 2 of the manual presents ln*R* as a "new metric for ecological meta-analysis" when in fact this metric is not new—it is simply the log transformation of the proportionate impact and has been used often as a response metric in ecological studies that predate *MetaWin*.[14–16] Calling ln*R* a "new" metric is indication of how entrenched *d* is in the meta-analytic arena and how seldom other metrics are considered.

Importantly, choosing a metric of effect size is a difficult task that involves conceptual issues that link the metric to the hypothesis, as well as statistical ones that require some knowledge of the properties of possible estimators of the desired quantity. We fear that many biologists performing meta-analyses will take the easy road and continue to use *d* because of its accessibility, without first evaluating their question and how best to quantify an effect in light of their question. At this point in time, it might have been safer had *MetaWin* been written only in the flexible mode, with no canned choice of the metric. This would require the user to ponder (at least briefly) the choice and applicability of an effect size metric, of which there are a multitude in biology. Thus far, however, there seems to be a misconception

Illustrating the importance of choosing the right metric



We used *MetaWin* to compare three classes of experiments and contrast three different metrics of effect size. The original papers reported effects of fish on snail density, and we classified each study based on the functional morphology of the fish species: "specialists" have highly modified pharyngeal muscles and bones suitable for shell crushing; "generalists" lack such structures and have more generalized diets; and "mixture" was assigned when the fish assemblage included both types and their separate effects could not be isolated. Details and references are given in Osenberg et al.[8] **Top:** Effect size defined using the standard metric, $d = (N_C – N_E/s)J$, where N_C and N_E are snail densities in the presence and absence of fish, s is the pooled standard deviation within treatment groups, and J is a correction for small sample bias.[5] **Middle:** Effect size defined as ln $(N_C \div N_E)$.[14] **Bottom:** Effect size defined as ln $(N_C \div N_E)/Pt$, where P is fish density and t is the duration of the experiment in days.[8] All analyses were conducted using a mixed model; results show pooled effect sizes (weighted means) and 95% confidence intervals. Notice that each metric returns similar qualitative results for the effects of generalists and specialists (i.e., specialists have larger—more deleterious—effects on snail dynamics than do generalists). However, the metrics diverge strikingly in the results for mixtures. Mixtures were comprised by only 20–50% (mean = 36%) specialists and thus should have demonstrated impacts only ~ 36% of that observed for specialists, given the negligible effect on generalists. This was only the case in the bottom panel, where the mixtures had effects that were 28% of those estimated for the specialists (vs. 146% and 93% for the other two metrics). Most importantly, the top panel does not yield a clear biological interpretation because the metric is defined primarily on statistical grounds. The middle panel summarizes results as proportionate impacts, but does not incorporate experimental duration. The bottom panel yields better estimates (smaller confidence intervals) because it accounts for among-study variation in both predator density and experimental duration. Further, the bottom panel is based on a metric explicitly linked to a biological model, in this case a plausible model of species interactions,[8] and is the only metric of the three that quantifies the impact as a change in a rate, specifically the per capita effect of fish on the instantaneous population growth rate (day^{-1}) of snails. These rate estimates can be readily incorporated into dynamic models of fish-snail interactions.

among ecologists and evolutionary biologists that meta-analysis means "using *d*" or statistically related metrics. This is unfortunate, and if the error persists, it will significantly impede the innovative application of meta-analytic tools. Just as there are infinite questions of interest to scientists, so too will there be an infinite number of metrics for meta-analysis. Choosing among them is the most important step in meta-analytic research, and the solution cannot be found by the blind application of tools developed by statisticians. Instead, measures of effect size must be derived from the theory and concepts that define each scientific discipline that uses meta-analysis (e.g., Box 1). The extent to which these steps are successfully implemented will be a testament to the future success of meta-analysis within each field. To the extent that they are overlooked, meta-analysis will fail to live up to its potential.

Null Hypothesis Tests

Most scientists are interested in quantifying effect sizes, determining the confidence in those estimates, and explaining the factors that might drive variation in effect among studies. Few are truly and primarily interested in combined tests of a null hypothesis of "no effect." If they were, many metrics (including *d*) would be suitable. Despite the growing recognition that null hypothesis tests are insufficient,[11–13] many statistical packages do not even provide estimates of the magnitude of treatment effects (and variances) as defaults when performing statistical tests (e.g., PROC GLM or ANOVA in SAS). *MetaWin*, in contrast, provides estimates of pooled effect sizes and their confidence intervals (using parametric and resampling techniques) in all analyses, in addition to the standard results from the null hypothesis tests of meta-analysis (e.g., "no effect overall," "no heterogeneity within classes," "no variation in effect size between classes"). This focus on the magnitude of effects and confidence intervals is encouraging, and suggests that *MetaWin*, and meta-analysis in general, will encourage more attention on estimation in primary studies. Any movement away from blind significance testing and the *P*-value culture[13] would be welcome. Just as the use of statistics in evolution and ecology originally focused on tests of null hypotheses and then expanded to more diverse applications, it is our hope that meta-analysis will also grow to address more complex questions that facilitate prediction, estimation, and the fitting of response surfaces.[17,18]

HOW APPLICATIONS IN COMPARATIVE BIOLOGY MIGHT ENRICH META-ANALYSIS

Practitioners of meta-analysis often argue that their tool could radically change the face of existing fields: e.g., "We predict that meta-analysis will have a substantial impact on the field of ecology."[6] We have no doubt that this is true—and reiterate Nelder's[19] argument that quantitative synthesis is a fundamental part of the scientific method, and as such should provide a radically different perspective than science conducted without quantitative synthesis! However, we think meta-analysis would go much further, be embraced by more scientists, and have a greater impact on science if it were viewed not so much as a tool to be imposed on existing fields, but as a tool to be transformed and defined by those fields. In that light, we believe comparative biology has much to bring to meta-analysis, and we conclude this essay by using comparative biology to illuminate our discussion of effect size metrics and null hypothesis tests.

Comparative biologists have been doing meta-analyses for many years (but without the appropriate jargon, or some of the established protocol). As a simple example, consider the many scaling relationships that have been examined by functional morphologists and physiologists, such as the relationship between metabolic rate (*R*) and body mass (*M*). In many of these studies, the goal is quite clear: e.g., to obtain quantitative estimates of parameters *a* and *b* in the function $R = aM^b$ and to quantify how these parameters vary among different groups of organisms. In some cases, the actual data consist of measures of oxygen in vessels with the organism present, and control vessels without the organism (to deal with other sources of oxygen depletion and generation). Such data are ripe for a classic meta-analysis. But note: 1) no comparative biologist would bother to test the null hypothesis that the metabolic rate was equal to zero (this would seem pointless and might raise the ire of the local Animal Care and Use Committee); 2) no physiologist would dare summarize the results from a single study by taking the difference between the mean oxygen concentrations and dividing by the pooled standard deviation; and 3) few scientists would bother with tests of the homogeneity of effect sizes—heterogeneity characterizes biological systems and provides the template that comparative biologists seek to quantify and explain—there is little need to test for its existence. Instead, the process of interest defines the appropriate metric (the difference in the rate of oxygen depletion, i.e., the metabolic rate of the study organism), the model defines the covariate (body size), and the question defines the comparisons to be made (e.g., contrasting flightless birds with their more flighted relatives).[20] Other examples of this approach include the scaling of filtering rates,[21] feeding performance,[22] and trophic level biomass (as a function of phosphorus loading).[23] In these examples, and throughout comparative biology, analyses are needed that explicitly incorporate covariates and examine functional relationships based on models that describe continuous variation in both the dependent and independent variables. Unfortunately, it is not yet possible to do regressions or use covariates in *MetaWin*. This will be a significant impediment to applications in comparative biology.

Given the need to address different problems, most of which will go beyond the capabilities of this first release of *MetaWin*, it seems likely many investigators will resort to standard statistical packages to analyze their meta-data sets. Under some conditions, this approach is relatively safe. However, there may be serious practical reasons to avoid such an approach.[1] For example, meta-data are likely to contain considerable heterogeneity due to variation in techniques, sample sizes, and classes of systems that comprise the data set. This heterogeneity is likely to lead to gross violations of the assumptions that underlie standard statistical tools.[1] Clearly, there is a vital need for more robust statistical tools—these would not only aid meta-analyses, but also primary analyses.

Finally, it is worth noting that many comparative approaches explicitly address the problems of non-independence, particularly arising from phylogenetic relationships. Indeed, realizing the full power of meta-analysis in evolutionary contexts will require that a phylogenetic perspective, e.g., be integrated into the analytic framework. A phylogenetic perspective simply helps us better define the population of theoretical studies to which the empirical results can be generalized, and thus is similar in its effects on meta-analysis to other sources of non-independence, e.g., due to publication bias or study bias. Techniques developed by comparative biologists to explore phylogenetic issues need to be brought to bear on the general problem of non-independence in meta-analysis.

Meta-analysis also has much to offer comparative biology. For example, most meta-analytic approaches explicitly incorporate estimates of error from single studies. This error arises from two sources: sampling (or within-study) error and systematic (or between-study) error. When sampling error is moderately large (relative to systematic error) and variable among studies, estimates should be weighted differentially (e.g., by the inverse of the variance of the estimator).[1] Weighting is typically not done in comparative analyses, although the precision of estimates can vary wildly. Instead, equal weight is usually given to each datum, independent of its precision. This likely leads to poor estimation of the parameters of interest to comparative biologists, and thus undermines the strong quantitative approach exemplified by many comparative studies.

Many of our comments in this essay reflect an underlying tension between biologically motivated questions and statistically motivated procedures. We believe that unless the question can be phrased in a clear biologically relevant context, specific and well-defined statistical metrics and approaches are useless. A good question aided by a metric that captures the process of interest can provide tremendous insight even if the statistical analyses are rather crude. As Tukey[24] said (in a line often cited by statisticians giving advice to ecologists who pride themselves on statistical rigor): "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." We have no doubt that meta-analysis will have a tremendous influence in biology—it already has—but its impact will be determined by the questions that can be addressed, which must be defined by the investigators, and not the details of the existing meta-analytic tools, which were defined by statisticians for other purposes in other disciplines. We urge the comparative biologists who use meta-analysis to help define those questions and the basic approaches needed to resolve those questions. The statistical refinements will naturally follow.

ACKNOWLEDGMENTS

C.W.O. thanks the National Center for Ecological Analysis and Synthesis, a center funded by NSF (DEB-94-21535), the University of California, Santa Barbara, and the State of California for its support of the working group *Meta-Analysis, Interaction Strength, and Effect Size: Application of Biological Models to the Synthesis of Experimental Data*. Many of the ideas reflected in this paper stem from discussions generated at the center with members of the working group. We also thank Peter Kareiva for helpful comments and for the invitation to write this essay.

REFERENCES

1 Hedges LV, Olkin I (1985)"Statistical Methods for Meta-Analysis." Orlando, FL: Academic Press.

2 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 133:62–67.

3 Glass GV (1976) Primary, secondary, and meta-analysis of research. Educ Res 5:3–8.

4 Rosenberg MS, Adams DC, Gurevitch J (1996) "*Meta Win*: Statistical Software for Meta-Analysis With Resampling Tests." Version 1.0. Sunderland, MA: Sinauer Associates.

5 Gurevitch J, Morrow LL, Wallace A, Walsh JS (1992) A meta-analysis of competition in field experiments. Am Nat 140:539–572.

6 Gurevitch J, Hedges LV (1993) Meta-analysis: Combining the results of independent experiments. In Scheiner SM, Gurevitch J (eds): "Design and Analysis of Ecological Experiments." New York: Chapman & Hall, pp 378–398.

7 Arnqvist G, Wooster D (1995) Meta-analysis—Synthesizing research findings in ecology and evolution. Trends Ecol Evol 10:236–240.

8 Osenberg CW, Sarnelle O, Cooper SD (1997) Effect size in ecological experiments: The application of biological models to meta-analysis. Am Nat 150: 798–812.

9 Finney DJ (1995) A statistician looks at meta-analysis. J Clin Epidemiol 48:87–103.

10 Petraitis PS (1998) How can we compare the importance of ecological processes if we never ask, "Compared to what?" In Resetarits W, Bernardo J (eds)."Conceptual Issues in Experimental Ecology." New York: Oxford University Press (in press).

11 Jones D, Matloff N (1986) Statistical hypothesis testing in biology: A contradiction in terms. J Econ Entomol 79:1156–1160.

12 Yoccoz NG (1991) Use, overuse and misuse of significance tests in evolutionary biology and ecology. Bull Ecol Soc Am 72:106–111.

13 Stewart-Oaten A (1996) Goals in environmental monitoring. In Schmitt RJ, Osenberg CW (eds): "Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats." San Diego: Academic Press, pp 17–27.

14 Cooper SD, Walde SJ, Peckarsky BL (1990) Prey exchange rates and the impact of predators on prey populations in streams. Ecology 71:1503–1514.

15 Sarnelle O (1992) Nutrient enrichment and grazer effects on phytoplankton in lakes. Ecology 74:551–560.

16 Schroeter SC, Dixon JD, Kastendiek J, Smith RO, Bence JR (1993) Detecting the ecological effects of environmental impacts: A case study of kelp forest invertebrates. Ecol Appl 3:331–350.

17 Li Y, Powers TE, Roth HD (1994) Random-effects linear regression meta-analysis models with application to the nitrogen dioxide health effects studies. J Air Waste Manage Assoc 44:261–270.

18 Vanhonacker WR (1996) Meta-analysis and response surface extrapolation: A least square approach. Am Stat 50:294–299.

19 Nelder JA (1986) Statistics, science and technology. JR Stat Soc A 149:109–121.

20 McNab BK (1994) Energy conservation and the evolution of flightlessness in birds. Am Nat 144:628–642.

21 Peters RH, Downing JA (1984) Empirical analysis of zooplankton filtering and feeding rates. Limnol Oceanogr 29:763–784.

22 Wainwright PC (1988) Morphology and ecology: Functional basis of feeding constraints in Caribbean labrid fishes. Ecology 69: 635–645.

23 Osenberg CW, Mittlebach GG (1996) The relative importance of resource limitation and predator limitation in food chains. In Polis GA, Winemiller GA (eds): "Food Webs: Integration of Patterns and Dynamics." New York: Chapman & Hall, pp 134–148.

24 Tukey JW (1962) The future of data analysis. Ann Math Stat 33:1–67.

Craig W. Osenberg
Colette M. St. Mary
Department of Zoology
University of Florida
Gainesville, FL 32611-8525
E-mail: osenberg@zoo.ufl.edu