

## RESOLVING ECOLOGICAL QUESTIONS THROUGH META-ANALYSIS: GOALS, METRICS, AND MODELS

CRAIG W. OSENBERG,<sup>1,5</sup> ORLANDO SARNELLE,<sup>2,6</sup> SCOTT D. COOPER,<sup>3</sup> AND ROBERT D. HOLT<sup>4</sup>

<sup>1</sup> Department of Zoology, University of Florida, Gainesville, Florida 32611-8525 USA

<sup>2</sup> National Center for Ecological Analysis and Synthesis, University of California,  
Santa Barbara, California 93101 USA

<sup>3</sup> Department of Ecology, Evolution, and Marine Biology, and Marine Science Institute,  
University of California, Santa Barbara, California 93106 USA

<sup>4</sup> Museum of Natural History, University of Kansas, Lawrence, Kansas 66045 USA

**Abstract.** We evaluate the goals of meta-analysis, critique its recent application in ecology, and highlight an approach that more explicitly links meta-analysis and ecological theory. One goal of meta-analysis is testing null hypotheses of no response to experimental manipulations. Many ecologists, however, are more interested in quantitatively measuring processes and examining their systematic variation across systems and conditions. This latter goal requires a suite of diverse, ecologically based metrics of effect size, with each appropriately matched to an ecological question of interest. By specifying ecological models, we can develop metrics of effect size that quantify the underlying process or response of interest and are insensitive to extraneous factors irrelevant to the focal question. A model will also help to delineate the set of studies that fit the question addressed by the meta-analysis.

We discuss factors that can give rise to heterogeneity in effect sizes (e.g., due to differences in experimental protocol, parameter values, or the structure of the models that describe system dynamics) and illustrate this variation using some simple models of plant competition. Variation in time scale will be one of the most common factors affecting a meta-analysis, by introducing heterogeneity in effect sizes. Different metrics will apply to different time scales, and time-series data will be vital in evaluating the appropriateness of different metrics to different collections of studies.

We then illustrate the application of ecological models, and associated metrics of effect size, in meta-analysis by discussing and/or synthesizing data on species interactions, mutual interference between consumers, and individual physiology. We also examine the use of metrics when no single, specific model applies to the synthesized studies. These examples illustrate that the diversity of ecological questions demands a diversity of ecologically meaningful metrics of effect size. The successful application of meta-analysis in ecology will benefit by clear and explicit linkages among ecological theory, the questions being addressed, and the metrics used to summarize the available information.

**Key words:** ecological models and meta-analysis; effect size in meta-analysis; interaction strength; meta-analysis goals, metrics, and models; metrics, ecological relevance; predation; ratio dependence; time scale.

### INTRODUCTION

For the past forty years, experiments have been heralded as the most powerful tool in the ecologist's field kit. Indeed, since the seminal work of Connell (1961), field experiments have revealed patterns in the dynamics of ecological systems, and the insights from these studies have provided a foundation for much of population and community ecology. As a result, ecologists have developed rigorous criteria for the design and

analysis of field experiments and have used the results of single studies as exemplars that define and mold ecological theory. Many important ecological hypotheses, however, cannot be tested using the standard single-study approach. Often what is needed is a comparative approach that examines how processes and responses vary across ecological systems. Literature reviews and the discussions of individual papers often attempt to put the results of disparate studies into a general ecological context. At present, however, these reviews remain largely qualitative and subjective, and certainly are not subjected to the same quantitative rigor as primary, experimental studies. Meta-analysis aims to rectify this shortcoming.

Manuscript received 11 November 1997; revised 2 July 1998; accepted 8 July 1998; final version received 14 September 1998. For reprints of this Special Feature, see footnote 1, p. 1103.

<sup>5</sup> E-mail: osenberg@zoo.ufl.edu

<sup>6</sup> Present address: Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan 48824-1222 USA.

Meta-analysis is the quantitative synthesis, analysis, and summary of a collection of studies (Hedges and Olkin 1985). Meta-analysis requires that the results of

each experiment be summarized with an estimate of the magnitude of the response to the manipulation (i.e., the "effect size"). In principle this response can be multivariate, although in practice it is univariate. Effect size, once extracted from each study, is the response that is subjected to further analysis. Although many issues related to meta-analysis have been discussed, and often hotly debated, in the ecological, medical, and social sciences (e.g., bias in reporting, the use of various statistical models, etc.; see Mann 1990, Gurevitch and Hedges 1993, Cooper and Hedges 1994, Arnquist and Wooster 1995; see also *Journal of Evolutionary Biology*, volume 10), there has been little discussion of the conceptual basis or implications of different metrics of effect size. Yet, defining the focal questions (and hence the metric of effect size) is the most fundamental task in conducting a meta-analysis (Osenberg et al. 1997, Osenberg and St. Mary 1998).

Three related but distinct goals underlie most meta-analyses and should influence the choice of a metric of effect size: (1) the construction of an aggregated and more powerful test of a null hypothesis, (2) the estimation of the magnitude of response (which might take the form of parameter estimation), and (3) the subsequent examination of the relationship between these estimates and various environmental and biological variables. Many of the earliest meta-analyses were focused on the first goal, a combined test of a hypothesis of "no effect." Aggregate tests are most instructive when each separate study yields equivocal results because of low statistical power. In such cases, a combined test provides a more powerful statistical test of the null hypothesis (e.g., Johnson et al. 1987, Vanderwerf 1992, Hechtel and Juliano 1997). Many investigators, however, have questioned the wisdom of focusing on tests of hypotheses of "no effect" (Jones and Matloff 1986, Yoccoz 1991, Stewart-Oaten 1996, Fernandez-Duque 1997). Indeed, it might be argued that any process studied repeatedly by ecologists probably has *some* effect, although the effect might be so small as to defy statistical detection in most experiments. For example, no one really doubts that competition, predation, or trophic cascades occur in natural systems. What does engender considerable interest and debate is the strength and relative importance of these phenomena across systems or environmental conditions (Quinn and Dunham 1983, Gurevitch et al. 1992, Sarnelle 1992, Osenberg and Mittelbach 1996, Steele 1997). As a consequence, the quantification of effect size and exploration of the pattern of variation in effect size among studies is a far more important goal of meta-analysis than the construction of powerful tests of null hypotheses (see also Stewart-Oaten 1996).

To facilitate the synthesis of ecological information and the exploration of general patterns within a body of ecological studies, it is critical that we match questions with appropriate metrics of effect size. Rather than seek a single or limited number of metrics, we

believe that each question might demand a conceptually distinct metric that is explicitly defined by the question and the ecological process(es) of interest. Importantly, quantifying the strength of a process (or the value of an ecological parameter) requires approaches distinct from those used to test null hypotheses. Unfortunately, metrics of effect size used in ecological meta-analyses are often derived from the methods used to test null hypotheses (see *Effect sizes commonly used in quantitative reviews*, below). Use of a narrow set of statistically motivated metrics potentially constrains the general inferences that can be drawn about patterns found across ecological studies (Osenberg et al. 1997, Osenberg and St. Mary 1998, Petraitis 1998). In this paper, we present a diversity of questions and suitable metrics to highlight this point, and illustrate the potential application of different metrics defined by reference to specific ecological questions and models. It is our hope that this exercise will greatly expand the future application of meta-analysis in resolving ecological questions and establish the effectiveness of meta-analysis in testing and refining ecological models and theory.

#### EFFECT SIZES COMMONLY USED IN QUANTITATIVE REVIEWS

Aggregate tests of a null hypothesis can be derived from a variety of statistical metrics obtained from each study, e.g.,  $P$  values (Fisher 1932), the original test statistics (e.g., the correlation coefficient,  $r$ , or the  $t$  value in a two-sample comparison, Rosenthal 1994), or the standardized difference between two treatments (e.g.,  $d$  in Hedges and Olkin 1985). In contrast, when the goal is to estimate the magnitude of responses, we need an alternative approach (Osenberg et al. 1997). Unfortunately, the preoccupation of many ecologists with statistically rigorous tests of null hypotheses has often led to confusion between biological and statistical significance (Yoccoz 1991, Stewart-Oaten 1996, Fernandez-Duque 1997). For example, some ecologists erroneously equate small  $P$  values (or large test statistics) with "large effects," and large  $P$  values (or small test statistics) with "small effects" or even the absence of an effect; e.g.,  $P > 0.05$  is often interpreted as affirming the validity of the null hypothesis. This error is well known, yet persists throughout the ecological literature, as exemplified by the use of "vote counting" (i.e., the tallying of significant and nonsignificant results) in many ecological syntheses (see Gurevitch et al. 1992, Gurevitch and Hedges 1999).

We illustrate problems with equating  $P$  values with biological significance by using data extracted from Peckarsky's (1985) study of the responses of different prey taxa to manipulations of the density of predatory stoneflies in streams. For each prey species, we calculated a simple measure of effect size (i.e., the log-transformed response ratio,  $\ln(N_E/N_C)$ , where  $N_E$  and  $N_C$  are the mean prey density with and without predators;

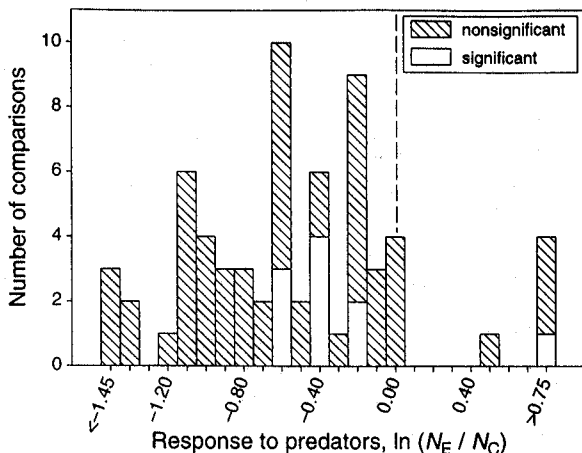


FIG. 1. Distribution of effects of stonefly predators on densities of different prey taxa in stream experiments. Results were classified by prey species that showed statistically significant ( $n = 10$ ) vs. nonsignificant ( $n = 56$ ) responses to predator manipulations. The response of each prey species to the predator manipulation was expressed using the log response ratio ( $\ln[N_E/N_C]$ , where  $N_E$  is the mean number of prey in the treatment with the predator and  $N_C$  is the mean number of prey in the treatment without the predator). Negative values indicate negative effects of the predator on prey density. Zero, indicated by the vertical dashed line, is the expected value when predators have no net effect on prey density. Data were taken from Table 4 in Peckarsky (1985).

Cooper et al. 1990, Peckarsky et al. 1990, Hedges et al. 1999). All data were taken from Table 4 in Peckarsky (1985) and separated into prey species that did ( $n = 10$ ) vs. those that did not ( $n = 56$ ) lead to rejection of the null hypothesis of "no effect" of predators on prey density (Fig. 1). Contrary to what one might expect, the distribution of effect sizes for prey that showed "nonsignificant" responses was not centered about zero (Fig. 1). Instead, the effects were predominantly negative. Most importantly, the mean effect of predators on prey showing "nonsignificant" responses was actually stronger than (although not statistically distinguishable from) the mean effect on prey showing "significant" responses (back-transformed means: 39% reduction in prey density [95% confidence interval, CI: 27–49%] vs. 17% reduction in density [CI: 37% augmentation–50% reduction]). Removing the one significant positive effect yielded an average effect (and confidence interval) for the "significant" responses that was even more comparable to the "nonsignificant" responses (mean and 95% CI: 33%, 26–41% reduction). Thus, the failure to reject the null hypothesis for the prey in the "nonsignificant" group appears to have resulted from a lack of power, rather than a reduced effect of predators on prey density. As this example illustrates,  $P$  values are an inappropriate measure of the biological magnitude of an effect. Similar arguments have been made against the use of test statistics (e.g., an  $F$  ratio or  $t$  value) and other statistically derived metrics (e.g.,  $d$ , the standardized difference;

Glass 1976, Cohen 1977) as indices of effect size (Finney 1995, Osenberg et al. 1997, Petraitis 1998).

#### CHOOSING A METRIC: THE NEED FOR BIOLOGICAL MODELS

A formal meta-analysis requires an estimate of effect size,  $e_i$ , and its variance,  $v(e_i)$ , from each of the  $i = 1, \dots, k$  studies that are used. There are no preconditions on the definition of  $e_i$ , and given the diversity of questions of interest to ecologists and evolutionary biologists, there is little reason to limit, a priori, the diversity of metrics of effect size. The metric that is applied in any particular meta-analysis, however, must have relevance to the question being addressed. Ecological models can provide an explicit framework for appropriately matching the question with a suitable metric, especially when the metric is represented as a parameter in the model. Models also help clarify the assumptions that underlie the derivation of a metric, and provide a clear basis for evaluating the contexts in which a particular metric is applicable (Osenberg et al. 1997, Downing et al. 1999).

#### Sources of variability among ecological studies

Theoretical investigations can help identify different kinds of variation, and thus help avoid (or correct) biases that may inadvertently contaminate particular metrics and associated analyses. Imagine that, for each system that has been studied, we have a dynamically sufficient (and preferably mechanistic) model that can account for variation in abundances over time, as well as predict with reasonable accuracy the effects of experimental perturbations (e.g., species removals). In such a case, we can envision at least four sources of variation that might distinguish studies and influence the choice of a metric of effect size.

**Level I: experimental variation.**—Ideally all studies in a meta-analysis would have been conducted in exactly the same way. In reality, experiments differ in various aspects that will influence the results (e.g., initial conditions, magnitude of perturbation, or duration of experimental manipulation). Even if different systems were described by the same dynamic model, with identical parameter values, these sources of variation could lead to systematic biases in parameter or effect-size estimation. Accounting for such biases will often require an explicit model of how the system functions (see *The right metric depends on the underlying dynamics*, below; see Downing et al. [1999]).

**Level II: parametric variation.**—Systems might be governed by the same basic dynamical model, yet differ in the values of model parameters. This can lead to variation among studies in the magnitude of response to any given treatment. Explaining such variation will often be the goal of a synthetic, quantitative review that uses meta-analysis. Theoretical studies can help elucidate how to minimize biases in estimating effects, particularly by removing experimental (Level I) vari-

ation so that parametric variation can be examined as a function of system traits (e.g., how the effect of competition varies across productivity gradients; see *The right metric depends on the underlying dynamics*, below).

*Level III: functional variation.*—Ecological systems might be sufficiently distinct that their dynamics cannot be accounted for only by parametric variation. Instead, the functions that describe the interactions between variables might assume different shapes. In many cases, we may not even know the specific algebraic functions describing a system's dynamics, but nonetheless have a qualitative sense of the cause-effect relationships that define the system's behavior. For example, consider a specialized pollinator-plant interaction that we are certain is mutually beneficial. In this case, an increase in the abundance of either species increases the fitness of the other. We may not be certain, however, whether the increase in the plant's fitness, for example, will be a linear, accelerating, or decelerating function of pollinator abundance.

*Level IV: structural variation.*—Finally, systems may differ in their causal relationships. Ecological theory and experiments have clearly shown that the dynamics of systems can be radically affected by the presence or absence of a single component (Leibold 1989, Abrams 1993, Holt 1997). For instance, some communities may have a predator influencing competitors, whereas such a predator might be absent from other communities. Changes in the abiotic environment may also change the qualitative nature of species interactions; e.g., benign commensal microorganisms can become lethal pathogens under some conditions. Although, in principle, a sophisticated model might be able to encompass all such variation by including a combination of parametric and functional variation, it is useful to separate variation in the qualitative causal structure of a system from the other kinds of variation noted above.

Recognizing these sources of variation helps us define strategies for conducting meta-analyses. Describing experimental (Level I) variation is seldom the final goal of a meta-analysis, but rather represents a preliminary or exploratory step focused on describing possible confounding influences on the results. Thus, meta-analyses often should include multiple stages of analysis, the first dealing specifically with experimental variation (e.g., resulting from time-scale issues; Osenberg et al. [1997], Downing et al. [1999]). Ideally, this is accomplished by knowing the model that governs the dynamics of the systems. Of course the correct model(s) is never known with certainty, so the challenge is to select the most appropriate model(s) given the goals of the study and the structure and dynamics of the focal systems, and to use the model(s) to guide the selection of a metric, define the conditions under which the metric is applicable, and delineate the empirical studies that can be synthesized using that metric.

We next illustrate why assumptions about the underlying dynamics are so critical when choosing a metric of effect size.

#### *The right metric depends on the underlying dynamics*

The choice of a metric should be guided by the question and process of interest, the structure and dynamics of the system being studied, and the effects of experimental treatments on the dynamics of the system. We illustrate the role of biological models in defining a suitable metric by reference to a simple, heuristic example that explores how the intensity of plant competition varies across productivity gradients (e.g., Goldberg et al. 1999). In our example, we focus on the specific problem of summarizing experiments that have examined the effects of competition from neighboring plants on the somatic growth of target plants. We assume that each experiment was set up with two treatments: a Control containing the ambient density of neighbors and a Removal in which all neighbors were removed. Each replicate had a central target individual of an initial mass that was similar across replicates and treatments. After  $t$  days the targets were harvested and their masses measured. Our aim, then, is to quantify the effect of competition using the mean individual mass at the start of the experiment ( $M_0$ , presumed equal between treatments), the mean individual mass at the end of the experiment in the presence and absence of competitors ( $M_{t,+}$ ,  $M_{t,-}$ , respectively), and the duration of the experiment ( $t$ ). Several metrics designed to quantify competitive effects have been suggested in the literature. These and related metrics include: Competitive Intensity, CI ( $=M_{t,-} - M_{t,+}$ ; Campbell and Grime 1992), Relative Competitive Intensity, RCI ( $= [M_{t,-} - M_{t,+}] / M_{t,-}$ ; e.g., Paine 1992, Wilson and Tilman 1993, Grace 1995, Goldberg et al. 1999), the Response Ratio, RR ( $= M_{t,+} / M_{t,-} = 1 - \text{RCI}$ ; Sarnelle 1992, Curtis and Wang 1998, Hedges et al. 1999), and the difference in per unit growth rates ( $\Delta r = \ln(M_{t,+} / M_{t,-}) / t = \ln(\text{RR}) / t$ ; Osenberg et al. 1997). Below we explore the behavior of these metrics under different assumptions about the dynamics of target growth and the effects of competition (i.e., functional, Level III, variation) to highlight how different dynamics require the application of different metrics of effect size (because of the way in which the metrics are affected by experimental variation).

Assume that growth of the target plant can be modeled as either an exponential or linear function of time defined by the growth rate,  $g$ :

$$M_t = M_0 e^{gt} \quad (1)$$

or

$$M_t = M_0 + gt \quad (2)$$

respectively, where  $g \geq 0$ . Note that the meaning (and units) of  $g$  varies between Eqs. 1 and 2. Further assume

that competition reduces the growth rate,  $g$ , in either an additive or multiplicative fashion, i.e.,

$$g = g_0 - c \quad 0 \leq c \leq g_0 \quad (3)$$

or

$$g = g_0/c \quad c \geq 1 \quad (4)$$

where  $g_0$  is the growth rate without competition, and  $c$  is the parameter that quantifies the competitive effect (larger values of  $c$  indicate more intense competition). The various combinations of these sets of equations describe four possible scenarios (Table 1). For example, combining Eqs. 1 and 3 produces a model analogous to the Lotka-Volterra equations for population dynamics. We re-emphasize that our intent here is not to model real systems, but rather to highlight links among model assumptions, metrics of effect size, and the inferences that might be drawn from the application of these metrics to systems with different dynamics.

We now ask how various metrics behave under the four different scenarios with the goal that the best metric isolates  $c$ , the effect of competition, from other sources of variation (e.g., experimental duration, and initial plant size). Thus, we ask how well different metrics isolate parametric (Level II) variation from experimental (Level I) variation, and how their performance varies depending upon assumptions about the system's dynamics. If all plants grow exponentially (Eq. 1) and competition affects growth additively (Eq. 3), then the Relative Competitive Intensity index (RCI) yields an effect size equal to  $1 - e^{-ct}$  (Table 1), i.e., RCI increases through time (from 0 to 1) as the size of target plants diverge in the two treatments. The rate at which RCI increases through time will be determined by the strength of competition,  $c$ . Thus, RCI is a function of the competitive effect,  $c$  (as it should be), as well as the duration of the study,  $t$ . As a result, two experiments, conducted in "identical" systems but lasting different durations, would yield different values of RCI, and thus lead to the erroneous inference that the intensity of competition varied between the systems. The response ratio (RR) is also a function of both the intensity of competition and experimental duration (Table 1), and Competitive Intensity (CI) is a function of the intensity of competition as well as experimental duration, initial plant mass, and plant growth in the absence of competition (Table 1). Thus, none of these three metrics exhibits the desirable property of isolating the competitive effect ( $c$ ) from the influence of extraneous factors that are likely to vary from one study to another (i.e., due to Level I variation, such as experimental duration). In the case of exponential growth and additive competitive effects, only  $\Delta r$  isolates the direct influence of competition from the other potentially confounding factors (Table 1, although the sign is reversed given the form of the response ratio).

If the dynamics of the systems are better described by one of the other models, then we obtain a different

result. In fact, for the other three scenarios (i.e., if plant growth is linear, or exponential with multiplicative effects of competition) none of the listed metrics (including  $\Delta r$ ) properly isolates  $c$  from the other parameters contained in the models (Table 1). In these cases, if commonly used metrics of competitive intensity were selected, then parametric (Level II) variation would always be confounded with experimental (Level I) variation. If each system that had been studied could be described by a common functional model, then the appropriate way to estimate  $c$  depends on the form of the model (i.e., the dynamics of the system; Table 1). Exponential growth requires a metric based on relative growth rates, whereas linear growth requires a metric based on absolute growth rates; additive competitive effects are estimated as a difference in individual growth rates (either measured on a per unit or absolute basis), whereas multiplicative competitive effects require the calculation of the ratio of growth rates. If each system is best described by a different model (i.e., due to Level III or IV variation), then there is no clear choice of a metric to use in a meta-analysis using all of the studies. These insights were not obvious until the possible metrics were matched to the possible models. Clearly, a metric should not be pulled randomly from a long list, or selected for mathematical or statistical convenience. The choice of metric can have serious effects on the results of a meta-analysis, and can clearly affect the interpretation of those results (Osenberg et al. 1997). For example, in the cases outlined above, several possible metrics were functions of individual plant growth in the absence of competition, which should vary positively with resource supply. As a consequence, an ill-chosen measure of competitive effect might change along an environmental gradient (e.g., a productivity gradient) for reasons completely independent of the strength of competition.

#### *Time-scale considerations: removing Level I variation*

One of the most important sources of Level I variation arises from variation in the length of time an experiment runs. An appropriate metric should be derived based on this consideration (Table 1). In many cases, however, the question being addressed (and hence the application of any given model and associated metric) may be relevant to only a particular range of time scales. Consider a population perturbed from a stable equilibrium. If the question being addressed concerns the initial effect of the perturbation on per capita growth rates, then the metric of choice is  $\Delta r$ ; e.g., if the control's rate of change is 0,  $\Delta r = \ln[N/N_0]/t$ , where  $N_0$  and  $N$  are the population densities immediately after the perturbation and  $t$  time units later, respectively (see *Examples . . . : Parameter estimation . . . : Interaction strength*, below). As  $t$  gets larger, feedbacks (e.g., indirect effects) may drive the system to a new equilibrium. As a result,  $\Delta r$  becomes smaller with

TABLE 1. Performance of various metrics in studies of plant competition, where target plants increase in mass according to different assumptions about the growth function (exponential [Eq. 1] or linear [Eq. 2]) and the effect of competition (additive [Eq. 3] or multiplicative [Eq. 4]).

| Metric†                     | Exponential growth (Eq. 1)  |  |
|-----------------------------|---|--|
|                             | Additive effect (Eq. 3)   | Multiplicative effect (Eq. 4)  |
| CI                          | $M_0 e^{g_0 t} (1 - e^{-ct})$   | $M_0 (e^{g_0 t} - e^{g_0 t/c})$  |
| RCI                         | $1 - e^{-ct}$   | $1 - e^{-g_0 t(1-1/c)}$  |
| RR                          | $e^{-ct}$   | $e^{-g_0 t(1-1/c)}$  |
| $\Delta r$                  | $-c$  | $-g_0(1 - 1/c)$  |
| Estimator of $c$            | $\ln(M_{t,-}/M_{t,+})/t \ddagger$   | $\ln(M_{t,-}/M_0)/\ln(M_{t,+}/M_0)$  |
| Interpretation of estimator | Difference in per unit growth rates,<br>$dM_{t,-}/M_{t,-} dt - dM_{t,+}/M_{t,+} dt$ | Ratio of per unit growth rates,<br>$(dM_{t,-}/M_{t,-} dt)/(dM_{t,+}/M_{t,+} dt)$ |

Notes:  $M_0$  is the mean individual mass at the start of the experiment;  $M_{t,+}$  and  $M_{t,-}$  are the mean individual masses at the end of the experiment in the presence and absence of competitors, respectively;  $t$  is the duration of the experiment,  $c$  is the effect of competition, and  $g_0$  is the growth constant without competition.

† CI = competitive intensity =  $M_{t,+} - M_{t,-}$ ; RCI = relative competitive intensity =  $(M_{t,-} - M_{t,+})/M_{t,-}$ ; RR = response ratio =  $M_{t,+}/M_{t,-} = 1 - \text{RCI}$ ;  $\Delta r$  = difference in relative growth rates (per gram) =  $\ln(M_{t,+}/M_{t,-})/t = \ln(\text{RR}/t)$ . The choice of the numerator and denominator of RR and  $\Delta r$  are somewhat arbitrary and can be converted to the alternative form by taking the reciprocal of RR or changing the sign of  $\Delta r$ .

‡  $c = -\Delta r$ .

time, asymptotically approaching zero. If studies differ in the length of time that treatments are applied, experiments that run longer (all else being equal) automatically will have a lower estimate for  $\Delta r$  (see Osenberg et al. [1997] and Downing et al. 1999 for empirical examples). In other cases (e.g., *Examples . . . : Beyond parameter estimation . . . : Long-term effects of grazers on algae*, below), alternative metrics (e.g., based on changes in equilibrium abundances, such as the difference in, or ratio of, densities between the two treatments) may be more appropriate than ones based on rates of change (e.g., see also Billick and Case 1994 and Osenberg et al. 1997).

Choosing the appropriate time scales for analyses based upon a particular metric of effect size is not straightforward. We see two ways to address the problems posed when the theoretical value of a metric varies through time. First, if we had a dynamic model for a particular system or class of systems, we could conduct numerical simulations to examine how serious these potential problems might be. Thus, one potential role of theory, which to our knowledge has rarely been used, is to delimit the temporal domain within which certain metrics should be used.

Second, we could take a more empirical approach. Time-series data from the systems being studied could be used to determine the time scales at which specific metrics are time invariant and actually estimate the process of interest (e.g., Osenberg et al. 1997 and Downing et al. 1999). Ideally, time-series data would be available for each study included in a meta-analysis. Unfortunately, most ecological field experiments are relatively short in duration, and often report responses only at the end of the experiment (perhaps with initial conditions). Without time series, models cannot be developed and validated, and the meta-analyst will need

to rely on experience and indirect assessments to evaluate the appropriateness of a metric and its underlying model.

In the absence of time series, it might be tempting to explore the relationship between effect size ( $e_i$ ) and duration ( $t_i$ ) across studies ( $i = 1, \dots, k$ ). We caution, however, that effect size and experimental duration might be correlated among studies, but that this correlation does not mean that the metric varies with time within a study. For example, when investigators anticipate strong (vs. subtle) effects, they might run experiments for shorter (vs. longer) durations, even if the metric of effect size is time invariant within each study. Because of the importance of time series for model validation in both meta-analytic and primary studies, we advocate increased collection and publication of time-series responses to experimental manipulations.

Explicit attention to time-scale considerations is essential to avoid confusion when the same term is used in different contexts. For example, Paine (1992) and Osenberg et al. (1997) both purport to measure "interaction strength"; however, Paine's metric is based on the assumption that the systems re-equilibrate following the experimental perturbation (Laska and Wootton 1998), whereas Osenberg et al.'s (1997) metric is based on the assumption that the systems' dynamics are transient and exponential. Extracting estimates of interaction strength from the literature and combining them in a meta-analysis would be problematic if both these definitions of interaction strength were used in the primary studies, because different types of interaction strength estimate different effects (e.g., direct vs. indirect effects) that operate on different time scales (e.g., short vs. long).

After Level I heterogeneity (e.g., due to time-scale effects) has been addressed, then the subsequent and

TABLE 1. Extended.

| Linear growth (Eq. 2)                                    |   |
|--|---|
| Additive effect (Eq. 3)                                  | Multiplicative effect (Eq. 4)                         |
| $ct$   | $g_0t(1 - 1/c)$                                       |
| $ct/(M_0 + g_0t)$  | $g_0t(1 - 1/c)/(M_0 + g_0t)$                          |
| $1 - [ct/(M_0 + g_0t)]$                                  | $(M_0 + g_0t/c)/(M_0 + g_0t)$                         |
| $\ln(1 - [ct/(M_0 + g_0t)])/t$                           | $\ln[(M_0 + g_0t/c)/(M_0 + g_0t)]/t$                  |
| $(M_{t,-} - M_{t,+})/t$                                  | $(M_{t,-} - M_0)/(M_{t,+} - M_0)$                     |
| Difference in absolute growth rates, $dM_-/dt - dM_+/dt$ | Ratio of absolute growth rates, $(dM_-/dt)/(dM_+/dt)$ |

main analyses may proceed. If a single model can be specified, then we often will be interested in exploring systematic variation in the response parameter or metric derived from the model. Examination of patterns in the variation of the metric can lead to the development of more-general models (e.g., where parameters now become functions of specific environmental or organismal traits), and to the creation of new hypotheses to be tested by subsequent experiments. If, on the other hand, a single, common, functional (and structural) model cannot be applied (e.g., if, in Table 1, some systems are governed by one set of dynamics, and some systems by other dynamics), then we might question whether the studies should be combined into a single analysis. Resolution of this question will depend on the extent to which results from systems governed by different dynamics can be interpreted within a common conceptual framework. We return to this issue of the implications of Level III and IV variation in the final section (*Beyond parameter estimation: Level III and IV variation*).

#### EXAMPLES OF BIOLOGICALLY-RELEVANT ESTIMATES OF EFFECT SIZE

##### *Parameter estimation: Level II variation*

Meta-analysis may be most useful in estimating parameters and analyzing patterns of variation in a given parameter across studies (i.e., exploring Level II variation while controlling for Level I variation). This goal has a long tradition in the sciences, dating back to P. R. Heyl's attempts in the 1930s to combine estimates of the gravitational constant (Hedges and Olkin 1985). Below, we provide five specific examples to highlight the diversity of potential applications of meta-analysis in ecology. In the first three examples, which deal with Level II variation, we estimate parameters and explore their variation among studies. In the final two examples, we address Level III and IV variation, where parameter estimation is not the goal because a single functional model could not be applied to all studies. In these two cases, conceptual models were still necessary to frame and answer the questions. In each of the five

examples we attempt to highlight a new insight drawn from the analysis that either clarifies the role of meta-analysis in synthesizing data and/or points to its possible limitations.

**Interaction strength.**—Many ecological studies purport to examine the effect of one species on another. Although definitions vary (Laska and Wootton 1998), we define interaction strength (between species  $j$  and  $i$ ) as the effect of an individual of species  $j$  on the per capita growth rates of species  $i$ , holding all other factors constant, i.e.,  $\partial(dN_i/N_i dt)/\partial N_j$ , where  $N_i$  and  $N_j$  are the densities of two interacting species. Under certain experimental conditions (see Osenberg et al. 1997, Laska and Wootton 1998), interaction strength can be estimated as

$$a_{ij} = \partial(dN_i/N_i dt)/\partial N_j \approx \Delta r/N_j \\ = \ln(N_{i,t,+}/N_{i,t,-})/tN_j \quad (5)$$

Estimating average interaction strength (and confidence intervals) among a collection of studies requires decisions about how to weight each individual estimate (Gurevitch and Hedges 1999). Typically, each estimate is weighted by the inverse of its variance (i.e., by its precision). In mixed-, or random-, effects models, this variance has two components: (1) within-study variance (i.e., experimental error), and (2) between-study variance (i.e., variance among studies in their true effect sizes; Gurevitch and Hedges 1999, Hedges et al. 1999). Partitioning of the total variance into the within- and between-study components can be useful in categorizing studies and developing predictive models.

Osenberg et al. (1997) used Eq. 5 to estimate the interaction strength ( $a_{ij}$ ) between fish (the predator) and gastropods (the prey) based on gastropod biomass in cages and ponds with and without fish. However, they performed unweighted analyses (i.e., they gave equal weight to all experiments). This approach prevented them from partitioning the variance components, and likely resulted in confidence intervals that were too large. Therefore, we re-analyzed their data using a mixed model (Rosenberg et al. 1997) with weighted meta-analytic procedures (Appendix A). Note that the estimator for  $a_{ij}$  (Eq. 6) is equivalent to a log response ratio (sensu Hedges et al. 1999) divided by  $tN_j$ . Thus, we estimated within-study variance using the expression in Hedges et al. (1999) for the log response ratio divided by  $t^2N_j^2$ . Because within-study replication and the number of studies were small, we also used the small-sample correction suggested by Hedges et al. (1999).

Based on the new, weighted analyses (in which among-study variances were fit separately for each group), fishes with a feeding morphology specialized for crushing snails (i.e., pumpkinseed, redear sunfish, and tench) had much larger effects on snail dynamics than did fishes with a generalized feeding morphology (i.e., bluegill sunfish, largemouth bass, and Eurasian

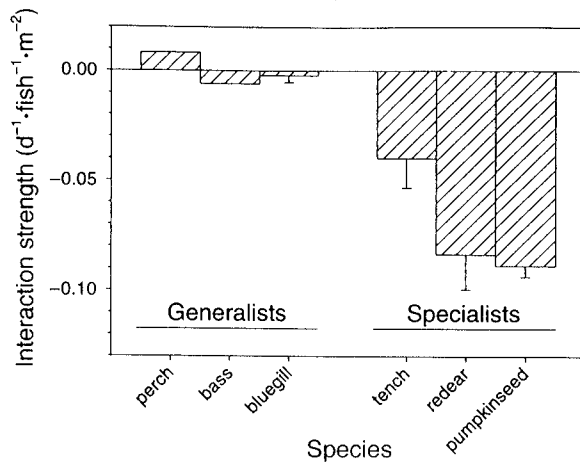


FIG. 2. Interaction strength (estimated using Eq. 6) between snails and six species of fish based on field experiments that manipulated fish density and recorded the response of snail biomass density after 43–93 d. Specialists (tench, redear, and pumpkinseed) have modified pharyngeal structures that enable the fish to crush snails. Generalists (Eurasian perch, largemouth bass, and bluegill) are not capable of crushing snails and therefore typically swallow them whole. Data were taken from Osenberg et al. (1997) and are provided in Appendix A. Weighted means and 95% confidence intervals (error bars are symmetrical) were based on a random-effects model that separately estimated among-study variances for each species of fish.

perch):  $a_{\text{snails, generalists}} = -0.0007$  (95% confidence interval, CI:  $-0.0038$ – $0.0025$ ),  $n = 9$ ;  $a_{\text{snails, specialists}} = -0.079$  (CI:  $-0.108$ – $-0.050$ ),  $n = 13$ . These weighted estimates actually agree well with the unweighted estimates reported by Osenberg et al. (1997). Importantly, however, the new analysis allowed us to assess the degree of heterogeneity in effect sizes among studies within a morphological class. There was no significant variation in effect size among studies within the generalist group (test of within-group homogeneity:  $Q_w = 3.88$ ,  $P = 0.87$ ); indeed, variation among studies accounted for 0% of the total variation (among-study plus mean within-study variance). In contrast, there was significant heterogeneity in effect sizes within the specialist group ( $Q_w = 199.0$ ,  $P < 0.0001$ ), with the among-study variance accounting for 75% of the total variance. We expected that much of the heterogeneity among studies within the specialist group could be attributed to variation in the effect of different fish species on snail density. Therefore, we redid the analysis, but classified effects by species (rather than morphological group). This greatly reduced the within-group heterogeneity, in part because pumpkinseed and redear had effects on snails that were more than twice as great as those of tench (Fig. 2). There was no significant heterogeneity among tench or redear studies ( $Q_w = 0.001$ ,  $P = 0.97$ , and  $Q_w = 4.99$ ,  $P = 0.17$ , respectively). Heterogeneity among pumpkinseed studies, however, remained significant ( $Q_w = 22.3$ ,  $P = 0.001$ ).

The among-study (but within-species) variance averaged only ~30% of the total variance for the three specialized species.

As more studies accumulate, it might be possible to explore the residual variation within a single species and relate this variation to environmental features, such as macrophyte density (Crowder and Cooper 1982), water temperature, or predator or prey body size (Osenberg and Mittelbach 1989) (see also Goldberg et al. [1999] and Englund et al. [1999] for other meta-analyses of species interactions). These estimates of interaction strength could be used in models of fish–snail dynamics to quantify the direct effects of an entire fish assemblage on snail dynamics across systems that differ in fish species composition, size-structure, and environmental characteristics.

**Mutual interference.**—The approach used above to quantify interaction strength assumes that the effects of fish on snails are linearly related to fish and snail densities. In many situations, however, there will be important nonlinearities in such interactions. These may arise from nonlinear functional responses produced by prey handling time, predator satiation, or predator interference. Mutual interference arises when predators reduce the instantaneous attack rate of other predators. Thus, as predator density increases, each predator has a decreasing effect on per capita prey survival (i.e., the attack coefficient is a decreasing function of predator density). Hassell and Varley (1969) proposed that the instantaneous attack rate,  $a$ , declined as a power function of predator density,  $P$  (e.g.,  $a(P) = \alpha P^{-m}$ ). Incorporating this into a Type II functional response yields

$$f(N, P) = \alpha NP^{-m} / (1 + \alpha t_h NP^{-m}) \quad (6)$$

where  $N$  and  $P$  are prey and predator density respectively,  $t_h$  is the handling time per prey,  $\alpha$  is the instantaneous attack rate of a solitary predator, and  $m$  measures the degree of mutual interference among predators (Arditi and Akçakaya 1990). Eq. 6 reduces to the standard prey-dependent Type II functional response when  $m = 0$  (i.e.,  $f$  is a function of only prey density,  $N$ , not  $P$ ) and to a pure ratio-dependent functional response when  $m = 1$  (i.e.,  $f$  is a function of the ratio of prey to predator densities,  $N/P$ ; Getz 1984, Arditi and Ginzburg 1989).

Arditi and Akçakaya (1990), noting that predator–prey dynamics are greatly affected by the value of  $m$ , surveyed the literature for data that could be used to estimate  $m$  and to evaluate if estimates were equal to 0 or 1, thereby distinguishing between prey-dependent and pure ratio-dependent functional responses. They found suitable data from 15 studies that reported attack rates of predators (and parasitoids) under variable densities of both predators and prey. They tested two null hypotheses with data from each study: that  $m_i$  (the estimate from study  $i$ :  $i = 1, \dots, 15$ ) was sampled from a population with true  $m = 0$  or true  $m = 1$ . They found



that 15 of 15 cases led to rejection of the hypothesis that  $m_i = 0$ , but only 3 of 15 cases led to rejection of  $m_i = 1$ . Given the general failure to reject the hypothesis that  $m = 1$  but the overwhelming evidence rejecting  $m = 0$ , they concluded that "it may therefore be acceptable to assume that  $m = 1$  in many of the systems analyzed here" and that the results provide "empirical support to the arguments . . . that ratio-dependent functional responses may be common . . ." (Arditi and Akçakaya 1990: 361).

We reanalyzed these data to highlight the application of meta-analysis in parameter estimation and the advantages of this approach over null-hypothesis testing. We took the 15 estimates of  $m$  and their variances from Table 2 in Arditi and Akçakaya (1990), and used a random-effects model (Rosenberg et al. 1997) to analyze the data (Appendix B). The pooled, weighted estimate of  $m$ ,  $m_w$ , was  $0.72 \pm 0.12$  (95% CI). Note that this confidence limit excludes both 0 and 1, suggesting that these predators do exhibit interference, but that it is weaker than assumed by pure ratio-dependent models. The collection of studies, however, was significantly heterogeneous ( $Q_w = 33.4$ ,  $P = 0.0025$ ), indicating that the studies were not each defined by the same value of  $m$ . Thus, although  $m_w < 1$ , some of the individual studies might have had  $m \geq 1$ . The meta-analytic results enable us to estimate what fraction of the studies had  $m \geq 1$ , given that we can specify the underlying distribution of true  $m$ 's. We assumed that the true  $m$ 's were normally distributed, with true mean = 0.72 (i.e., the estimated weighted mean) and true among-study variance = 0.0263 (i.e., the estimated among-study variance). Assuming that this distribution describes the larger population from which the 15 studies were sampled, then only 4.3% of these systems have  $m \geq 1$ .

It is important to note that these 15 studies were not randomly drawn from all possible predator-prey systems, but probably represent systems in which the original investigators suspected interference was occurring (hence, they manipulated predator density). Thus, based on this probable bias and the meta-analytic results, we conclude, in contrast to Arditi and Akçakaya (1990), that pure ratio-dependent functional responses are probably *not* common. Even among the systems most likely to exhibit ratio dependence, approximately 95% have interference parameters  $< 1$ . Obviously, nature is not black and white, with all systems falling into either the "prey-dependent" or "ratio-dependent" schemes. Instead, effects are often continuously distributed, with many intermediates. Meta-analysis is a powerful tool to analyze and describe this variation, although care must be taken in generalizing the analytical results beyond the studies being summarized. Given the inherent problems arising from study and publication bias in meta-analysis, the issue of generalization requires further study (Gurevitch and Hedges

1999; P. Petraitis and A. Dunham, *personal communication*).

*Mechanistic, physiologically structured models.*— Individual-based models are often parameterized by mining the literature for estimates of feeding, respiration, egestion, survival, and reproductive rates (e.g., Gurney et al. 1990, McCauley et al. 1990). Typically, representative estimates are culled, using informal procedures, from the many estimates of a parameter that are found in the literature. In contrast, meta-analytic procedures could be used to formally combine these estimates (to obtain weighted means, confidence intervals, and among-study variances). Replacing informal protocols with more explicit meta-analytic techniques should improve estimation procedures. Importantly, estimates of confidence intervals and among-study variances from the meta-analysis could be used to bracket the range of variation in the parameter (as illustrated in the section *Mutual interference*, above) and thus permit more informed sensitivity analyses that vary the parameter values. In some cases, the results could also be used to establish the variation in individual-based parameters to include within a single model.

Mechanistic approaches (including individual-based models) also might be useful in furthering the development of general ecological models (Dunham and Beaupre 1998). For example, Kooijman (1993), Müller and Nisbet (1997), and E. B. Müller and R. M. Nisbet (*unpublished manuscript*) developed a mechanistic dynamic energy-budget model based on rules of energy acquisition, energy allocation, and the uptake of toxicants. Under constant food conditions, this model simplifies to a form of the von Bertalanffy growth equation:

$$L_t = L_x - (L_x - L_0)e^{-\gamma t} \quad (7)$$

where

$$L_x = L_{x,0}/(1 + c/K)^2 \quad (8)$$

$$\gamma = \gamma_0(1 + c/K) \quad (9)$$

and where  $L_t$  is the length of the organism at time  $t$ ,  $L_0$  is its initial length,  $L_x$  is the maximum attainable length,  $\gamma$  is the growth constant,  $L_{x,0}$  and  $\gamma_0$  are the values in the absence of toxicants,  $c$  is the concentration of the toxicant, and  $K$  is a half-saturation constant that quantifies the sensitivity of the organism's growth to the toxicant. Smaller values of  $K$  indicate greater sensitivity to the toxicant. E. B. Müller and R. M. Nisbet (*unpublished manuscript*) estimated  $K$  using data for two species of mussels transplanted to several sites that varied in exposure to a wastewater discharge that contained elevated concentrations of barium (Higashi et al. 1992, Osenberg et al. 1992). Barium accumulated in the mussel shells and served as an index of exposure to the wastewater. Their analysis yielded  $K = 186.4 \pm 54.5$  nmol Ba/g shell for *Mytilus edulis* and  $K = 82.7 \pm 7.3$  nmol/g for *M. californianus*. Thus, *M. californianus*

*nianus* was more sensitive than *M. edulis* to wastewater exposure. There are vast amounts of data on the growth of bivalves in polluted and non-polluted environments; thus, meta-analysis (using  $K$  as a measure of effect size) could provide a useful tool to summarize these data and evaluate how sensitivity varies across species (as in this example), different types or mixtures of toxicants, or different environments (e.g., with different food levels or temperatures). Because the model and associated measure of sensitivity (i.e.,  $K$ ) are based on specific allocation and toxicological mechanisms, these meta-analytic results could improve our knowledge of the responses of organisms to pollutants.

*Beyond parameter estimation:  
Level III and IV variation*

Often, it may not be possible to specify an appropriate model or to point to a particular parameter of interest in analyzing and interpreting the results of ecological investigations. In such cases, other approaches can be used, but they still require explicit discussion of the processes or concepts of interest. We illustrate these situations with two cases, resource limitation and grazer–algal interactions, which are examined in detail elsewhere (Sarnelle 1992, Osenberg and Mittelbach 1996, Downing et al. 1999).

**Resource limitation.** The magnitude of resource limitation for consumers has implications for population dynamics, species interactions, the rates of evolution of feeding traits and energy-allocation strategies, and ecosystem processes (e.g., Power 1992, Osenberg and Mittelbach 1996). Examination of resource limitation has been impeded by the lack of a single, operational definition. Here, we define “resource limitation” as the extent to which the per capita production of a population (or aggregated unit, such as trophic level) is reduced due to low availability of resources (Osenberg and Mittelbach 1996). For example, if we consider a gradient of resource availability,  $R$ , over which all other factors are held constant, we expect per capita population growth ( $dN/Ndt$ ) to increase monotonically to an asymptote (i.e., as  $R \rightarrow \infty$ ,  $dN/Ndt \rightarrow r_m$  because further increases are not possible due to physiological limits or because other resources become limiting). Under ambient resources, the population grows at a per capita rate,  $r$ . The difference between  $r$  and  $r_m$  is a measure of limitation. Limitation can be assessed empirically by estimating the difference in the per capita growth rate of the consumer population under ambient and enriched conditions, assuming that enriched conditions have saturated resource consumption, that the direct effects of enrichment dominate the divergence of the consumer populations under the two treatments (i.e., responses are measured before indirect feedbacks arise), and that the consumer population is of the same size in the two treatments at the start of the experiment (see also Downing et al. 1999). In such cases, data must be restricted to responses measured over short time

scales, so as not to confound measures of limitation with feedbacks that dampen the response of consumers to increased resources (e.g., due to increases in the consumer’s predators, or decreases in the availability of other limiting resources).

Notice that this approach does not demand specification of the exact form of the relationship between per capita production and resource availability, but rather relies only on its qualitative shape (monotonic) and is thus an example where Level III variation can exist among the studies. Downing et al. (1999) applied this approach to examine patterns of nutrient limitation for marine phytoplankton, and Osenberg and Mittelbach (1996) used this approach to quantify the relative effects of resource and predator limitation in an aquatic food chain.

**Long-term effects of grazers on algae.** As a final example, which deals with Level IV variation, Sarnelle (1992) used a meta-analytical approach to examine competing hypotheses (i.e., model structures) offered to explain the strong increase in phytoplankton biomass with nutrient enrichment in lakes. He started with two simple grazer–algae (predator–prey) models. One model assumed that the algae were a homogeneous set of species that were equally vulnerable to grazing. The other assumed that the algae were available in two forms that differed in their susceptibilities to grazing. The models diverged in their predictions about how equilibrium algal biomass would respond to grazer manipulations across systems that varied in algal carrying capacity (i.e., as determined by phosphorus levels). Sarnelle calculated the ratio of algal biomass in low-grazer vs. high-grazer treatments in field experiments and examined the relationship between this ratio and total phosphorus concentration in the water column. The model with only edible algae predicted a linear relationship between the response ratio and total phosphorus, whereas the model with both edible and inedible algae predicted various nonlinear relationships. The observed linearity of the relationship between the response ratio and total phosphorus suggested that zooplanktivorous fish, rather than the structure of the algal assemblage, played a major role in driving the increase in phytoplankton biomass with enrichment, a conclusion with important applied implications (Shapiro and Wright 1984, Newman 1993) that has since been confirmed by a larger survey of non-experimental data (Mazumder 1994).

Two aspects of Sarnelle’s study provide general lessons for meta-analysis. First, because the predictions of the models and the effect-size metric were based on equilibrium responses, it was critical to restrict the data used in the meta-analysis to those experiments that had lasted long enough to re-equilibrate. This restriction is exactly opposite to that imposed in the previous resource-limitation example, which reinforces our contention that the choice of metric and the selection of

data should be driven by an explicit formulation of the question being addressed.

Secondly, Sarnelle (1992) included whole-lake experiments in his synthesis because they are the only experiments conducted on the exact spatial scale implied by the central question (i.e., why does phytoplankton biomass increase with enrichment in lakes?). Whole-lake experiments are usually unreplicated and so must be excluded from meta-analyses that require weighted analyses or that use metrics of effect size (such as  $d$ ) that are functions of within-study variance (e.g., see Wooster 1994, Brett and Goldman 1996, Curtis 1996 for restrictions based on these criteria). Meta-analyses that exclude whole-system studies (or other studies that lack estimates of within-study variance) run the risk of yielding results that are biased by enclosure artifacts and other problems inherent to experiments conducted on small spatial scales (Cooper et al. 1998). More precise estimates do not always correspond to more accurate ones. On the other hand, the absence of within-study variances prevented Sarnelle (1992) from performing weighted analyses and estimating the among-system variance (see *Interaction strength* and *Mutual interference*, above, for examples of the advantages associated with weighted analyses; also see Gurevitch and Hedges 1999).

#### CONCLUSIONS AND CAVEATS

Innovative quantitative synthesis and comparative analyses are not new to ecology or evolutionary biology. Comparative biologists and comparative limnologists have been synthesizing large data sets for decades (e.g., Damuth 1981, Peters 1983, Downing and Peters 1984, Watson et al. 1992). Often these data sets are derived from non-experimental studies. Although we have emphasized the application of meta-analysis to experimental data, meta-analysis is just as applicable to non-experimental data. In dealing with experimental or observational data, the choice of a model and metric will influence the results of, and hence the inferences drawn from, a meta-analysis. It is critical, therefore, that specification of the model be an explicit part of the meta-analytic protocol as well as an integral part of the primary studies. This is particularly important in meta-analyses using experimental data because suitable measures of effect size often are not presented in the primary literature, but must instead be calculated from the reported data. These derived metrics of effect often are not presented in primary studies because the raw data are sufficient to perform null-hypothesis tests, which are the principal goal of many experimental studies. This contrasts markedly with comparative biology, where the question often is focused more on parameter estimation (e.g., scaling relationships, Peters 1983) rather than null-hypothesis tests. Because effects are not explicitly quantified, it is often incumbent upon the meta-analyst to evaluate the appropriateness of the data to the questions.

Despite the theme of our paper, the application of biological models and the use of ecologically relevant metrics is not sufficient to guarantee sound inference. A variety of other issues, primarily statistical, must also be considered (e.g., Gurevitch and Hedges 1999, Hedges et al. 1999). Two issues deserve particular attention. First, once the question has been defined and an appropriate metric chosen, statisticians can play a critical role in developing efficient and unbiased estimators of these metrics and specifying the statistical circumstances under which the estimators are best applied (Hedges et al. 1999). The dynamic interplay between ecological concepts and the application of statistical tools is critical to the success of primary investigations as well as meta-analyses and will be most successful if we remember that the answers we obtain from the statistical analyses are (1) at best approximate (Tukey 1962), and (2) only as good as the ecological ideas that motivated the analysis. Rigorous application of well-understood statistics to vague ecological questions and poor data will do little to advance the field.

Second, even the most thorough and careful meta-analysis will contain bias. Systems are not chosen randomly for study, nor are results published without regard to their conclusions. Single papers often contain multiple experiments, single experiments often yield multiple comparisons, and single investigators vary tremendously in their productivity, methodology, and proclivity to study particular systems and questions. As a result, there will always be uncertainty about the population to which inferences apply. Through careful thought and innovative analyses, however, we can begin to confront these sources of bias (Gurevitch and Hedges 1999). Synthesis is a critical part of the scientific method, and ecology will benefit by developing tools (such as meta-analysis) that enhance our syntheses by making them more quantitative, more explicit, and better integrated with the theoretical issues that we wish to resolve.

#### ACKNOWLEDGMENTS

This work was conducted as part of the Meta-analysis Working Group ("*Meta-analysis, interaction strength and effect size: application of biological models to the synthesis of experimental data*") supported by the National Center for Ecological Analysis and Synthesis (NCEAS), a Center funded by NSF (DEB-94-21535), the University of California-Santa Barbara, and the State of California. Additional support also was provided for O. Sarnelle as a NCEAS Postdoctoral Associate, and through grants from the Minerals Management Service (U.S. Department of Interior, 14-35-0001-30761 to C. W. Osenberg) and NSF (DEB-9528445 to C. W. Osenberg, DEB93-08375 and DEB96-29473 to O. Sarnelle, and DEB-9407591 to S. D. Cooper. Our ideas and presentation have benefitted immensely from discussions with the entire Working Group and by the support provided by the staff of NCEAS. We also thank J. Wilson for assistance with the project, C. St. Mary, J. Dykes, E. Müller, and W. Rice for helpful discussions, and D. Goldberg and two anonymous reviewers for helpful comments on a previous draft. We also are grateful to D. Goldberg for her contributions to the entire Special Feature.

## LITERATURE CITED

- Abrams, P. 1993. Effects of increased productivity on the abundance of trophic levels. *American Naturalist* **141**:351–371.
- Arditi, R., and H. R. Akçakaya. 1990. Underestimation of mutual interference of predators. *Oecologia* **83**:358–361.
- Arditi, R., and L. R. Ginzburg. 1989. Coupling in predator-prey dynamics: ratio-dependence. *Journal of Theoretical Biology* **139**:311–326.
- Arnqvist, G., and D. Wooster. 1995. Meta-analysis—synthesizing research findings in ecology and evolution. *Trends in Ecology and Evolution* **10**:236–240.
- Billick, I., and T. J. Case. 1994. Higher order interactions in ecological communities: what are they and how can they be detected? *Ecology* **75**:1529–1543.
- Brett, M. T., and C. Goldman. 1996. A meta-analysis of the freshwater trophic cascade. *Proceedings of the National Academy of Sciences* **93**:7723–7726.
- Campbell, B. D., and J. P. Grime. 1992. An experimental test of plant strategy theory. *Ecology* **73**:15–29.
- Cohen, J. 1977. *Statistical power analysis for the behavioral sciences*. Academic Press, New York, New York, USA.
- Connell, J. H. 1961. The influence of interspecific competition and other factors on the distribution of the barnacle, *Chthamalus stellatus*. *Ecology* **42**:710–723.
- Cooper, H., and L. V. Hedges, editors. 1994. *The handbook of research synthesis*. Russell Sage Foundation, New York, New York, USA.
- Cooper, S. D., S. Diehl, K. Kratz, and O. Sarnelle. 1998. Implications of scale for patterns and processes in stream ecology. *Australian Journal of Ecology* **23**:27–40.
- Cooper, S. D., S. J. Walde, and B. L. Peckarsky. 1990. Prey exchange rates and the impact of predators on prey populations in streams. *Ecology* **71**:1503–1514.
- Crowder, L. B., and W. E. Cooper. 1982. Habitat structural complexity and the interaction between bluegills and their prey. *Ecology* **63**:1802–1813.
- Curtis, P. S. 1996. A meta-analysis of leaf gas exchange and nitrogen in trees grown under elevated carbon dioxide. *Plant, Cell and Environment* **19**:127–137.
- Curtis, P. S., and X. Wang. 1998. A meta-analysis of elevated CO<sub>2</sub> effects on woody plant mass, form, and physiology. *Oecologia* **113**:299–313.
- Damuth, J. 1981. Population density and body size in mammals. *Nature* **290**:699–700.
- 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.
- Downing, J. A., and R. H. Peters. 1984. Empirical analysis of zooplankton filtering and feeding rates. *Limnology and Oceanography* **29**:763–784.
- Dunham, A., and S. J. Beaupre. 1998. Ecological experiments: scale, phenomenology, mechanism, and the illusion of generality. Pages 27–49 in W. J. Resetarits, Jr., and J. Bernardo, editors. *Experimental ecology: issues and perspectives*. Oxford University Press, New York, New York, USA.
- Englund, G., O. Sarnelle, and S. D. Cooper. 1999. The importance of data-selection criteria: meta-analyses of stream predation experiments. *Ecology* **80**:1132–1141.
- Fernandez-Duque, E. 1997. Comparing and combining data across studies: alternatives to significance testing. *Oikos* **79**:616–618.
- Finney, D. J. 1995. A statistician looks at meta-analysis. *Journal of Clinical Epidemiology* **48**:87–103.
- Fisher, R. F. 1932. *Statistical methods for research workers*. Fourth edition. Oliver and Boyd, London, UK.
- Getz, W. 1984. Population dynamics: a per capita resource approach. *Journal of Theoretical Biology* **108**:623–643.
- Glass, G. V. 1976. Primary, secondary, and meta-analysis of research. *Educational Researcher* **5**:3–8.
- Goldberg, D. E., T. Rajaniemi, J. Gurevitch, and A. Stewart-Oaten. 1999. Empirical approaches to quantifying interaction intensity: competition and facilitation along productivity gradients. *Ecology* **80**:1118–1131.
- Grace, J. B. 1995. On the measurement of plant competition intensity. *Ecology* **76**:305–308.
- Gurevitch, J., and L. V. Hedges. 1993. Meta-analysis: combining the results of independent experiments. Pages 378–398 in S. M. Scheiner and J. Gurevitch, editors. *Design and analysis of ecological experiments*. Chapman & Hall, New York, New York, USA.
- Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. *Ecology* **80**:1142–1149.
- Gurevitch, J., L. L. Morrow, A. Wallace, and J. S. Walsh. 1992. A meta-analysis of competition in field experiments. *American Naturalist* **140**:539–572.
- Gurney, W. S. C., E. McCauley, R. M. Nisbet, and W. W. Murdoch. 1990. The physiological ecology of *Daphnia*: a dynamic model of growth and reproduction. *Ecology* **71**:716–732.
- Hassell, M. P., and G. C. Varley. 1969. New inductive population model for insect parasites and its bearing on biological control. *Nature* **223**:1133–1137.
- Hechtel, L. J., and S. A. Juliano. 1997. Effects of a predator on prey metamorphosis: plastic responses by prey or selective mortality? *Ecology* **78**:838–851.
- Hedges, L. V., J. Gurevitch, and P. Curtis. 1999. The meta-analysis of response ratios in experimental ecology. *Ecology* **80**:1150–1156.
- Hedges, L. V., and I. Olkin. 1985. *Statistical methods for meta-analysis*. Academic Press, Orlando, Florida USA.
- Higashi, R. M., G. N. Cherr, C. A. Bergens, and T. W.-M. Fan. 1992. An approach to toxicant isolation from a produced water source in the Santa Barbara channel. Pages 223–233 in J. P. Ray and F. R. Englehardt, editors. *Produced water: technological/environmental issues and solutions*. Plenum, New York, New York, USA.
- Holt, R. D. 1997. Community modules. Pages 333–350 in A. C. Gange and V. K. Brown, editors. *Multitrophic interactions in terrestrial systems*. Blackwell Science, Oxford, UK.
- Kooijman, S. A. L. M. 1993. *Dynamic energy budgets in biological systems*. Cambridge University Press, Cambridge, UK.
- Johnson, D. M., C. L. Pierce, T. H. Martin, C. N. Watson, R. E. Bohanan, and P. H. Crowley. 1987. Prey depletion by odonate larvae: combining evidence from multiple field experiments. *Ecology* **68**:1459–1465.
- Jones, D., and N. Matloff. 1986. Statistical hypothesis testing in biology: a contradiction in terms. *Journal of Economic Entomology* **79**:1156–1160.
- Laska, M. S., and J. T. Wootton. 1998. Theoretical concepts and empirical approaches to measuring interaction strength. *Ecology* **79**:461–476.
- Leibold, M. A. 1989. Resource edibility and the effects of predators and productivity on the outcome of trophic interactions. *American Naturalist* **134**:922–949.
- Mann, C. 1990. Meta-analysis in the breach. *Science* **275**:384–386.
- Mazumder, A. 1994. Patterns of algal biomass in dominant odd- vs. even-link lake ecosystems. *Ecology* **75**:1141–1149.
- McCauley, E., W. W. Murdoch, R. M. Nisbet, and W. S. C. Gurney. 1990. The physiological ecology of *Daphnia*: development of a model of growth and reproduction. *Ecology* **71**:703–715.
- Müller, E. B., and R. M. Nisbet. 1997. Modeling the effect of toxicants on the parameters of dynamic energy budget

- models. Pages 71–81 in F. J. Dwyer, T. R. Doane, and M. L. Hinman, editors. *Environmental toxicology and risk assessment. Volume 6. modeling and risk assessment.* American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.
- Newman, E. I. 1993. *Applied ecology.* Blackwell, Oxford, UK.
- Osenberg, C. W., and G. G. Mittelbach. 1989. The effects of body size on the predator–prey interaction between pumpkinseed sunfish and gastropods. *Ecological Monographs* **59**:405–432.
- Osenberg, C. W., and G. G. Mittelbach. 1996. The relative importance of resource limitation and predator limitation in food chains. Pages 134–148 in G. A. Polis and K. O. Winemiller, editors. *Food webs: integration of patterns and dynamics.* Chapman & Hall, New York, New York, USA.
- Osenberg, C. W., and C. M. St. Mary. 1998. Meta-analysis: synthesis or statistical subjugation? *Integrative Biology: Issues, News and Views* **1**:43–48.
- Osenberg, C. W., O. Sarnelle, and S. D. Cooper. 1997. Effect size in ecological experiments: the application of biological models to meta-analysis. *American Naturalist* **150**:798–812.
- Osenberg, C. W., R. J. Schmitt, S. J. Holbrook, and D. Canestro. 1992. Spatial scale of ecological effects associated with an open coast discharge of produced water. Pages 387–402 in J. P. Ray and F. R. Englehardt, editors. *Produced water: technological/environmental issues and solutions.* Plenum, New York, New York, USA.
- Paine, R. T. 1992. Food-web analysis through field measurement of per capita interaction strength. *Nature* **355**:73–75.
- Peckarsky, B. L. 1985. Do predaceous stoneflies and siltation affect the structure of stream insect communities colonizing enclosures? *Canadian Journal of Zoology* **63**:1519–1530.
- Peckarsky, B. L., S. C. Horn, and B. Statzner. 1990. Stonefly predation along a hydraulic gradient: a field test of the harsh–benign hypothesis. *Freshwater Biology* **24**:181–191.
- Peters, R. H. 1983. *The ecological implications of body size.* Cambridge University Press, Cambridge, UK.
- Petraitis, P. S. 1998. How can we compare the importance of ecological processes if we never ask, “compared to what?” Pages 183–201 in W. J. Reseraris, Jr., and J. Bernardo, editors. *Experimental ecology: issues and perspectives.* Oxford University Press, New York, New York, USA.
- Power, M. 1992. Top-down and bottom-up forces in food webs: do plants have primacy? *Ecology* **73**:733–746.
- Quinn, J. F., and A. E. Dunham. 1983. On hypothesis testing in ecology and evolution. *American Naturalist* **122**:602–617.
- 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. 1994. Parametric measures of effect size. Pages 231–244 in H. Cooper and L. V. Hedges, editors. *The handbook of research synthesis.* Russell Sage Foundation, New York, New York, USA.
- Sarnelle, O. 1992. Nutrient enrichment and grazer effects on phytoplankton in lakes. *Ecology* **74**:551–560.
- Shapiro, J., and D. I. Wright. 1984. Lake restoration by biomanipulation. Round Lake, Minnesota—the first two years. *Freshwater Biology* **14**:371–383.
- Steele, M. A. 1997. The relative importance of processes affecting recruitment of two temperate reef fishes. *Ecology* **78**:129–145.
- Stewart-Oaten, A. 1996. Goals in environmental monitoring. Pages 17–27 in R. J. Schmitt and C. W. Osenberg, editors. *Detecting ecological impacts: concepts and applications in coastal habitats.* Academic Press, San Diego, California, USA.
- Tukey, J. W. 1962. The future of data analysis. *Annals of Mathematical Statistics* **33**:1–67.
- Vanderwerf, E. 1992. Lack's clutch size hypothesis: an examination of the evidence using meta-analysis. *Ecology* **73**:1699–1705.
- Watson, S., E. McCauley, and J. A. Downing. 1992. Sigmoid relationships between phosphorus, algal biomass, and algal community structure. *Canadian Journal of Fisheries and Aquatic Sciences* **49**:2605–2610.
- Wilson, S. D., and D. Tilman. 1993. Plant competition and resource availability in response to disturbance and fertilization. *Ecology* **74**:599–611.
- Wooster, D. 1994. Predator impacts on stream benthic prey. *Oecologia* **99**:7–15.
- Yoccoz, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* **72**:106–111.

#### APPENDIX A

A table reporting fish effects on snail biomass density is available in ESA's Electronic Data Archive: *Ecological Archives* E080-005-S1.

#### APPENDIX B

A table reporting mutual interference parameters, modified from Arditi and Akçakaya (1991) is available in ESA's Electronic Data Archive: *Ecological Archives* E080-005-S2.