Ecology, 91(9), 2010, pp. 2561–2564

Achieving synthesis with meta-analysis by combining and comparing all available studies

Marc J. Lajeunesse[1]

National Evolutionary Synthesis Center, 2024 West Main Street, Suite A200, Durham, North Carolina 27705-4667 USA

Introduction

Advances in ecology do not require meta-analysis to answer questions. Yet, what meta-analysis provides is an opportunity to explore why multiple independent tests of these questions can have different outcomes. This exploration is a way meta-analysis can achieve synthesis: by identifying explanations for variation in research while isolating which concepts are applicable over a wide variety of contexts (Glass 1976, Greenland 1994). However, Whittaker (2010) argues—perhaps with some justification based on an audit of multiple conflicting reviews—that a more strict approach to meta-analysis would be more useful for ecology. Specifically, he favors a "best evidence synthesis" that combines both quantitative and qualitative reviewing techniques to answer more narrow questions with only high quality studies (following Slavin 1986, 1994).

My intention with this commentary is to explore how stringent the inclusion criteria should be for meta-analysis and the consequences for the breadth or narrowness of the resulting review. I primarily focus on two issues that have received little attention in ecological meta-analysis. First, how the intention and purpose of meta-analysis can impact the scope of the review, and second, how different philosophies on quality assessment can shape the inferences obtained from such reviews. To make these points, I rely heavily on previous discussions from the medical and social sciences about the application of the narrow "best evidence" approach over the broad exploratory alternative. For example, the best evidence approach qualitatively assesses study quality prior to synthesis; whereas the exploratory approach evaluates quality empirically (see Thompson 1994, Eysenck 1995). These opposing philosophies on how quality is treated can significantly alter the rewards of synthesis, perhaps resulting in a review with too few studies to make any useful generalization or a review that lacks the precision to estimate a biologically meaningful effect (van der Velde et al. 2007). Progress in ecological meta-analysis need not develop in isolation from advances in the medical or social sciences, and I hope that by briefly engaging Whittaker's argument for more narrow reviews with the literature of these fields, I can identify how meta-analysis can be used to validate ecological theory, and why to achieve this goal it is necessary to be broad and inclusive of all available research.

Manuscript received 21 August 2009; revised 20 November 2009; accepted 7 December 2009. Corresponding Editor: D. R. Strong. For reprints of this Forum, see footnote 1, p. 2534.

[1] E-mail: marc.lajeunesse@NESCent.org

Defining the Scope of the Review

Whittaker (2010) argues that a broad scope for meta-analysis is too inclusive and that answering narrow questions with a select group of studies is the only useful approach for synthesizing research (following Slavin 1986). However, when the scope of the review is defined this way, it has an explicit goal: to estimate as accurately as possible a specific, critical parameter of interest, for instance, a point estimate (average) of the overall shape of published species-productivity curves. This goal assumes that the overall research outcome can only be estimated from studies deemed consistent (homogeneous) by the reviewer. Otherwise, including a broad mix of studies might bring into question the validity of the overall effect. This lack of stringent inclusion criteria is what Whittaker concludes as the "mega-mistake" of previous meta-analyses on species-productivity relationships. Their scope was too broad and their results were too imprecise to validate theory.

However, why should the scope of meta-analysis focus solely on the precise estimation of pooled research outcomes? Precise point estimates are useful for parameterizing models or calculating the statistical power of future experiments (that is, only when effect sizes are the unit of the review). Yet such applications of meta-analytical results rarely if ever get used in subsequent primary research (Cooper et al. 2005). Point estimates paired with confidence intervals of effect sizes are also important to evaluate non-zero results when studies are weighted by sampling precision as in traditional meta-analysis (Hedges and Olkin 1985). But when studies are treated equally statistically (as in unweighted analyses), or when there are too few studies to synthesize, then the likelihood of making a review-level error is high (see Lajeunesse and Forbes 2003). Further, if the purpose of meta-analysis is to provide a more precise portrayal of an ecological phenomenon,

then the findings of studies should never be treated equally. This is because large within-study sampling error can influence the over- or under-estimation of a biological effect when results are pooled across few studies (Jüni et al. 1999). This focus on point estimates and lack of weighting clearly has influenced the results of meta-analyses on species-productivity curves—given the sensitivity of hypothesis tests and the variation in pooled results when studies are included/excluded from a given review (see Hillebrand and Cardinale 2010, Whittaker 2010).

Perhaps exploring what factors contribute to variation in research across a broad pool of studies would be more rewarding and effective to validating ecological theory (Anello and Fleiss 1995, Gøtzsche 2000). An important criterion for synthesis is validation through convergent confirmation of independent research using a diversity of experimental designs and measurements (Campbell and Fiske 1959, Strauss and Smith 2009). Given that ecological phenomena are likely multi-characteristic, multi-method processes, then restricting the scope of the review to studies with similar designs and measurements can only provide a narrow view of the biological effect of interest. Further, when heterogeneous results that define multiple operations of the same ecological construct are combined and compared, then something essential is learned about this biological effect beyond what each operation captures individually (Hall et al. 1994). This "triangulation" of the ecological phenomenon is what a reviewer achieves when they paint an inclusive picture of the literature (sensu Glass 1976), and when they are concerned with a wide range of questions regardless of the nature in design and quality of studies reviewed. Pooling research based on a combination of methodologies also insures that the variance of the ecological process reflects this process and not any one methodological artifact (Strauss and Smith 2009). Cleary, ecological theory will prove robust if it is applicable over a diversity of research.

Reviewers need to anticipate this heterogeneity across ecological studies, and embrace it as an opportunity to explore variation and to test hypotheses. Having a broad scope for meta-analysis demands that the review reconcile differences between studies with dissimilar results: this can lead to an enriched explanation of the research problem (Glasziou and Sanders 2002). For example, in seeking explanations of divergent results, the reviewer may uncover unexpected results or unseen factors moderating biological effects (I further elaborate on moderator variables in *Eligibility criteria and quality assessment*; also see Greenland 1994). These novel relationships can serve as stepping points for future experiments.

Eligibility Criteria and Quality Assessment

Slavin (1986) proposed the "best evidence" approach for meta-analysis because expert opinion, which is the predominant form of study inclusion of qualitative (narrative) reviews, is almost abandoned or at least underemphasized in quantitative reviews. Slavin argued that expert opinion was still necessary for meta-analysis; otherwise, how would a meta-analyst exclude the "garbage" from their review and prevent erroneous conclusions based on the inclusion of these data? Here strict eligibility (inclusion/exclusion) criteria serve as the reviewer's sieve for sorting the quality of research, leaving only the "best evidence" to review.

Whittaker (2010) revisits these issues, and proposes a fairly rigorous set of eligibility criteria for studies on species-productivity curves. Again, a "best evidence" synthesis requires detailed criteria to filter studies and to create a homogeneous data set. It is understandable why standardized selection criteria would be useful because (1) these types of quality judgments can be subjective and need clear guidelines (see Jørgensen et al. 2006); (2) inter-reviewer agreement on quality is low (Verhagen et al. 2001); and (3) clearly reported and uniform criteria is a way to improve the repeatability of results from multiple independent reviews of the same population of studies (Jadad et al. 1997, Hopayian 2001, Stroupa et al. 2001, Pullin and Stewart 2006, Peinemann et al. 2008). A lack of a common protocol appears systemic for meta-analyses on species–productivity relationships, where differences in quality judgments and data extraction among different research groups resulted in poorly matching data sets for the same research domain (Ellison 2010).

However, when eligibility criteria prune a population of 63 studies to four (see Whittaker 2010), then there is serious need to evaluate what exactly the "best evidence" approach achieves. Erroneous elimination of a prohibitive number of studies is not a solution to handling variation due to study "quality." Would it not be a greater service to the field to empirically address and test the relevance of these issues regarding quality as defined by the selection criteria? That is, to gather all the studies relevant to the conceptual topic under study, and then empirically test whether these differences (i.e., any factor presumably affecting quality) actually influence research outcomes. For example, contrasting the findings from groups of studies with and without these problems, or through sensitivity analyses where collections of studies are excluded from the overall synthesis to evaluate their weight on the pooled conclusions (Thompson 1994). Should a meta-analysis detect a difference between these groups, then (1) this provides practical information for future experiments to avoid these problems, (2) there is a solid rationale for why these studies should be included or excluded from the overall review, and (3) more sophisticated approaches such as statistics based on meta-regression techniques (analogous to an analysis of covariance) can be use to integrate issues on quality into the overall analysis (see Thompson and Higgins 2002).

An exploratory meta-analysis emphasizes evidence over opinion and seeks to provide a synthesis that is independent from reviewer bias in addition to more

subtle problems due to within-study sampling error. Homogeneity statistics have been explicitly developed for meta-analysis to evaluate whether variation exists across studies beyond the predicted sampling error, and whether studies should be pooled or grouped among moderator effects (Hedges and Olkin 1985). These moderators or predicted dimensions where studies fail to be "perfect" can be tested empirically, and then this evidence can be used as justification for a more narrow review or at least shape the eligibility criteria of future meta-analyses (Lau et al. 1998). In addition, homogeneity statistics evaluate whether these moderators make a difference when pooling studies and whether the causal relationship across these moderator groups is obtained despite their differences (Song et al. 2001). This approach (as well as meta-regression) allows for cross checking for internal consistency or reliability within a collection of studies deemed poor quality, while also retaining the important advantage of maintaining external validity of the ecological theory when results are pooled across methods (see *Defining the scope of the review*; Strauss and Smith 2009).

Blending and integrating a variety of data and methods also avoids errors introduced by expert opinion that can lead to biased (nonrandom) data sets. For example, a reviewer may formulate criteria based on a study they perceive as a "gold standard" for evidence because it found strong positive effects. However, sampling error alone can generate strong positive effects, and the efficiency meta-analytical statistics to account for this source of bias requires that data sets form a nonrandom sample of the population (Rosenthal 1991). Yet publication bias and taxonomic bias are already mechanisms that generate non-random data sets for ecological meta-analysis: there is no need to further exacerbate these problems by having strict selection criteria. These potential sources of bias in the population of studies available for review is why issues on quality should be explored with meta-analysis rather than used as a rationale for excluding research a priori before synthesis.

Conclusions

I believe the advantage of mixing a broad pool of research is clear: it allows for the systematic evaluation of factors that can explain variation in research, while simultaneously providing a complete summary of the current standing of a research domain (Gøtzsche 2000). However to date, there has not yet been any strong philosophical objection to having a broad scope for meta-analysis in ecology as weathered in the social and medical sciences—given the nearly geometric uptake of ecological meta-analysis since its introduction by Gurevitch et al. (1992). But what should be gleaned from Whittaker's critique is that there is a continued need for discussion about the function and purpose of meta-analysis for ecology. In addition, there are many issues unique to ecological meta-analysis that remain unaddressed; such as, a lack of effect size metrics that quantify the outcomes of more complicated experimental designs beyond the typical control–treatment contrast, and methods that account for the non-independence among effect size data (see Lajeunesse 2009).

Discussion on these issues would clarify what standards of the review process should be used as best practices, and what guidelines are necessary to improve inferences of reviews and the quality of meta-analyses (Jadad et al. 1997, Moher et al. 1999). Other statistical fields in biology have benefited tremendously from similar discussion. Debates on applications of the comparative phylogenetic method have since stabilized to where it is now uncommon to compare characteristics of multiple species without considering information on their shared evolutionary history (see Garland et al. 2005).

I anticipate that future discussion on ecological meta-analysis will stabilize to the following protocol: (1) eligibility criteria are broad and inclusive but fully reported in reviews; (2) studies are not treated equally when pooling results and are weighted by an estimate of study precision (e.g., sampling error); (3) sensitivity analyses, moderator groupings, and meta-regression are used to evaluate and integrate issues on quality and design of studies; (4) biological effects of interest are then evaluated using similar methods (testing conceptual hypotheses is inappropriate until methodological biases are considered first); (5) publication bias and other factors known to generate nonrandom data sets are explored to provide justification that the observed pooled effect is unbiased evidence for the ecological process of interest; and finally, (6) the "best evidence" synthesis is used as a heuristic tool only after a global synthesis of all available studies to test specific hypotheses, extract effect sizes for model parameterization, or the prognostic calculation of statistical power for future experiments.

Acknowledgments

I thank Aaron Ellison and two anonymous reviewers for valuable feedback on this manuscript. Funding for the work on this comment was provided by an NSF grant to the National Evolutionary Synthesis Center (EF-0423641).

Literature Cited

Anello, C., and J. L. Fleiss. 1995. Exploratory or analytic meta-analysis: Should we distinguish between them? Journal of Clinical Epidemiology 48:109–116.

Campbell, D. T., and D. W. Fiske. 1959. Convergent and discriminate validation by the multitrait-multimethod matrix. Psychological Bulletin 56:81–105.

Cooper, N. J., D. R. Jones, and A. J. Sutton. 2005. The use of systematic reviews when designing studies. Clinical Trials 2: 260–264.

Ellison, A. M. 2010. Repeatability and transparency in ecological research. Ecology 91:2536–2539.

Eysenck, H. J. 1995. Meta-analysis or best-evidence synthesis? Journal of Evaluation in Clinical Practice 1:29–36.

Garland, T., Jr., A. F. Bennett, and E. L. Rezende. 2005. Phylogenetic approaches in comparative physiology. Journal of Experimental Biology 208:3015–3035.

Glass, G. V. 1976. Primary, secondary, and meta-analysis. Educational Researcher 5:3–8.
Glasziou, P. P., and S. L. Sanders. 2002. Investigating causes of heterogeneity in systematic reviews. Statistics in Medicine 21: 1503–1511.
Gøtzsche, P. C. 2000. Why we need a broad perspective on meta-analysis: it may be crucially important for patients. BMJ 321:585–586.
Greenland, S. 1994. Invited commentary: a critical look at some popular meta-analytic methods. American Journal of Epidemiology 140:290–296.
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.
Hall, J. A., L. Tickle-Degnen, R. Rosenthal, and F. Mosteller. 1994. Hypothesis and problems in research synthesis. Pages 17–28 *in* L. V. Hedges and H. Cooper, editors. The handbook of research synthesis. Russell Sage Foundation, New York, New York, USA.
Hedges, L. V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, Orlando, Florida, USA.
Hillebrand, H., and B. J. Cardinale. 2010. A critique for meta-analyses and the productivity–diversity relationship. Ecology 91:2545–2549.
Hopayian, K. 2001. The need for caution in interpreting high quality systematic reviews. BMJ 323:681–684.
Jadad, A. R., D. J. Cook, and G. P. Browman. 1997. A guide to interpreting discordant systematic reviews. Canadian Medical Association Journal 156:1411–1416.
Jørgensen, A. W., J. Hilden, and P. C. Gøtzsche. 2006. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. BMJ 333:782.
Jüni, P., A. Witschi, R. Bloch, and M. Egger. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. Journal of the American Medical Association 282:1054–1060.
Lajeunesse, M. J. 2009. Meta-analysis and the comparative phylogenetic method. American Naturalist 174:369–381.
Lajeunesse, M. J., and M. R. Forbes. 2003. Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques. Ecology Letters 6:448–454.
Lau, J., J. P. A. Ioannidis, and C. H. Schmid. 1998. Summing up evidence: one answer is not always enough. Lancet 351: 123–127.
Moher, D., D. Cook, S. Eastwood, I. Olkin, D. Rennie, and D. F. Stroup. 1999. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Lancet 354:1896–1900.
Peinemann, F., N. McGauran, S. Sauerland, and S. Lange. 2008. Disagreement in primary study selection between systematic reviews on negative pressure wound therapy. BMC Medical Research Methodology 8:41.
Pullin, A. S., and G. B. Stewart. 2006. Guidelines for systematic review in conservation and environmental management. Conservation Biology 20:1647–1656.
Rosenthal, R. 1991. Meta-analytic procedures for social research. Sage, Newbury Park, California, USA.
Slavin, R. E. 1986. Best evidence synthesis: an alternative to meta-analytic and traditional reviews. Educational Researcher 15:5–11.
Slavin, R. E. 1994. Best evidence synthesis: an intelligent alternative to meta-analysis. Journal of Clinical Epidemiology 48:9–18.
Song, F., T. A. Sheldon, A. J. Sutton, K. R. Abrams, and D. R. Jones. 2001. Methods for exploring heterogeneity in meta-analysis. Evaluation and the Health Professions 24: 26–151.
Strauss, M. E., and G. T. Smith. 2009. Construct validity: advances in theory and methodology. Annual Review of Clinical Psychology 5:1–25.
Stroupa, D. F., S. B. Thackera, C. M. Olsonb, R. M. Glassc, and L. Hutwagnera. 2001. Characteristics of meta-analyses related to acceptance for publication in a medical journal. Journal of Clinical Epidemiology 54:655–660.
Thompson, S. G. 1994. Why sources of heterogeneity in meta-analysis should be investigated. British Medical Journal 309: 1351–1355.
Thompson, S. G., and J. P. T. Higgins. 2002. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 21:1559–1573.
van der Velde, G., M. van Tulder, P. Côté, S. Hogg-Johnson, P. Aker, and J. D. Cassidy. 2007. The sensitivity of review results to methods used to appraise and incorporate trial quality into data synthesis. Spine 32:796–806.
Verhagen, A. P., H. C. W de Vet, R. A. de Bie, M. Boers, and P. A. van den Brandt. 2001. The art of quality assessment of RCTs included in systematic reviews. Journal of Clinical Epidemiology 54:651–654.
Whittaker, R. J. 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91:2522–2533.