CONCEPTS & SYNTHESIS
EMPHASIZING NEW IDEAS TO STIMULATE RESEARCH IN ECOLOGY

*Ecology*, 91(9), 2010, pp. 2522–2533

# Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship

Robert J. Whittaker[1]

*Biodiversity Research Group, Oxford University Centre for the Environment, South Parks Road, Oxford OX1 3QY United Kingdom*

*Abstract*. The form of the species richness–productivity relationship (SRPR) is both theoretically important and contentious. In an effort to distill general patterns, ecologists have undertaken meta-analyses, within which each SRPR data set is first classified into one of five alternative forms: positive, humped (unimodal), negative, U-shaped (unimodal), and no relationship. Herein, I first provide a critique of this approach, based on 68 plant data sets/studies used in three meta-analyses published in *Ecology*. The meta-analyses are shown to have resulted in highly divergent outcomes, inconsistent and often highly inappropriate classification of data sets, and the introduction and multiplication of errors from one meta-analysis to the next. I therefore call on the ecological community at large to adopt a far more rigorous and critical attitude to the use of meta-analysis. Second, I develop the argument that the literature on the SRPR continues to be bedeviled by a common failing to appreciate the fundamental importance of the scale of analysis, beginning with the confusion evident between concepts of grain, focus, and extent. I postulate that variation in the form of the SRPR at fine scales of analysis owes much to artifacts of the sampling regime adopted. An improved understanding may emerge from combining sampling theory with an understanding of the factors controlling the form of species abundance distributions and species accumulation curves.

*Key words: confounding variables; diversity theory; meta-analysis; plants; primary productivity; scale.*

## Introduction

Species richness and productivity are two fundamental properties of (plant) ecological systems and the relationship between them has long been a subject of interest (e.g., Pianka 1966, Odum 1969). In experimental analyses using small plots the focus has sometimes been on how changing species richness changes system net primary productivity, but more usually the relationships is viewed, as herein, from the perspective of species richness as the dependent variable. The question that arose and which is at issue in the present paper is: what is the form of the species richness–productivity relationship (SRPR)? Is it (1) humped (unimodal), (2) U-shaped (negative unimodal), (3) positive monotonic, (4) negative monotonic, or is there (5) no relationship describable (i.e., neither linear nor unimodal)? The question is being asked because it is arguably fundamental to a mechanistic understanding of ecological diversity patterns (Whittaker et al. 2001) and because the relationship is poorly understood and contentious. The publication of a major meta-analysis of the SRPR including 121 plant data sets (90 of which are terrestrial systems, the rest aquatic) by Mittelbach et al. (2001) initially appeared to make an important contribution to understanding this problem, but closer examination revealed serious failings, leading Whittaker and Heegaard (2003) to call for the meta-analysis to be redone at consistent scales of analysis using more rigorous data-gathering and analytical protocols. I now realize that this call was a mistake on our part, because the data and protocols do not appear to exist to allow meaningful meta-analysis (cf. Slavin 1995). Three meta-analyses later, I now call for an end to meta-analyses of the SRPR, and a profound change in the criteria apparently being used by those undertaking, and reviewing submitted meta-analyses in ecology.

Subsequent to our critique and an accompanying defense by Mittelbach et al. (2003), Gillman and Wright (2006) responded to the challenge and reran a full meta-

Manuscript received 23 May 2008; revised 11 September 2008; accepted 8 December 2008. Corresponding Editor: D. R. Strong.

[1] E-mail: robert.whittaker@ouce.ox.ac.uk

analysis for plants (terrestrial systems), adding a further 37 studies to those previously gathered by Mittelbach et al. (2001). Their analysis endorsed all the criticisms leveled by Whittaker and Heegaard (2003) and contrary to the claims of Mittelbach et al. (2003) that the original analysis was robust, obtained substantially different results. Gillman and Wright's (2006) paper is in substance a worthy and critical reanalysis, and it is thus with some regret that I note below errors of detail in their paper. Mittelbach et al. (2001) has turned out to be a significant paper, attracting over 300 citations thus far (ISI data), with remarkably few to date noting the existence of the two critical reanalyses (Table 1) or that the paper may be an unreliable analysis. Meanwhile, a third meta-analysis of the SRPR for plants, by Pärtel et al. (2007), has now been published, like each of the foregoing papers, in *Ecology*. Pärtel et al. (2007) claimed to build directly on the Mittelbach et al. (2001) data base, did not refer at all to Whittaker and Heegaard (2003) and side-stepped Gillman and Wright's (2006) damning reanalysis with a single "but see." As I show below, if you do take the trouble to "go and see," what you find is that none of the meta-analyses agree with one another on how to classify a large proportion of the data sets in their analyses, raising immediate concerns over the approach and doubts as to whether they constitute repeatable science.

The meta-analysis approach is supposed to provide an objective means of summing up the emergent outcome of numerous tests of the same thing (e.g., the effectiveness of a new medicine or treatment) by compiling the results of previously published analyses and objectively analyzing the distribution of the outcomes (Slavin 1995). Unfortunately, in many areas of ecology, sampling system and design properties are virtually unique from study to study, and potentially confounding factors abound. Moreover, the aims of the original studies have often been profoundly different from those of the meta-analyses, providing some form of data that may be scavenged and recycled, but not necessarily that are fit for purpose. Such problems affect other areas of science, including medicine (Slavin 1995), but I suspect may be particularly acute in ecology. In recent work on the SRPR, this has meant that some form of original analysis of the data sets (or something approximating the original data) has had to be undertaken case-by-case prior to assessing the emergent outcomes. The authors of the meta-analysis are not therefore objectively assessing objective tests of the SRPR made by *previous* authors: rather they are themselves undertaking extensive primary analyses in order first to classify each study before compiling the findings for meta-analysis.

Undertaking such analysis and interpreting the outcome requires careful exposition and discussion. Within the meta-analyses, perhaps as a result of journal restrictions on pagination, next to no space is given to the data properties of the source papers, appropriateness of the analyses, or contextualization of the end result (e.g., scrutinize Partël et al. 2007). It appears, moreover, that the "meta-" part of the analysis overwhelms the usual critical instincts of reviewers and readers who fail to dig into the underlying original case analyses. This has enabled, as I show here (see Appendix A) the passage of regrettable and often elementary sequences of errors, compounded from one meta-analysis to the next.

Table 1. Citations to the papers by Mittelbach et al. (2001), Whittaker and Heegaard (2003), and Gillman and Wright (2006), respectively M2001, WH2003, and GW2006, between 2003 and May 2008 inclusive, according to a search using ISI Web of Science on 9 September 2008.

| Year | M2001 | WH2003 | GW2006 |
|---|---|---|---|
| 2003 | 50 | 1 | NA |
| 2004 | 45 | 2 | NA |
| 2005 | 53 | 14 | NA |
| 2006 | 57 | 10 | 1 |
| 2007 | 59 | 9 | 9 |
| [2008] | [30] | [5] | [7] |

*Notes:* Data for 2008 are given in square brackets as the year is incomplete. "NA" indicates not applicable.

All three of the meta-analyses contain error, although this is least apparent in the worthy attempt by Wright and Gillman (2006) to re-do using clear, stated criteria, the analysis for plants. The paper by Pärtel et al. (2007) provides no stated criteria or methods for the classification of studies and it attributes SRPR form inconsistently and inappropriately, often to fundamentally inadmissible data sets. These errors are repeated and compounded in a subsequent paper by the same team based on the same classification of studies (Laanisto et al. 2008), upon which I make little direct comment other than to regret its publication.

The primary purpose of the present paper is to ask that we draw a line under the whole approach. I anticipate that there may be some form of rejoinder published to this paper defending the meta-analytical approach, and so against that eventuality I ask that you, the reader, take the final call on whether the SRPR meta-analyses can be relied upon as repeatable ecological science. Colleagues, before you next cite their findings, please take the time to read this critical re-evaluation, and read at least some of the source papers (e.g., Beadle 1966, Flenley 1969, Wheeler and Giller 1982, Ehrman and Cocks 1990, Williams et al. 1996; and the following three papers of E. M. O'Brien's, which in fact present the same data set: O'Brien 1993, 1998, and O'Brien et al. 1998) in the light of the evidence and commentary I present in Appendix A. Having done so, why not set a class exercise for your students to read a few each, and then run an evaluation exercise with agreed criteria as to the attribution of SRPR form? See what *you* find. I predict that you will not wish to rely further upon the findings or meta-data presented in these meta-analyses and that it will lead into a wider discussion of the role of such analyses generally in ecology. My second goal is to develop the argument that the form of the SRPR is intrinsically scale-dependent and that much of the variation apparent at small focal scales of analysis constitutes an artifact of the use of inadequate plot sizes and protocols in the source literature.

## Criteria for Inclusion of a Study in a Meta-Analysis of the SRPR (for Plants)

The start point for any meta-analysis has to be to establish a set of protocols for searching out case studies, criteria for including/excluding them, and adopting, a priori, a particular analytical strategy, statistical approach and probability level (Slavin 1995). Here I comment only on the criteria for including/excluding a data set (for discussion of the other issues see Whittaker and Heegaard 2003, Mittelbach et al. 2003, Gillman and Wright 2006). While Mittelbach et al. (2001) analyzed both plant and animal SRPR, the focus of the other meta-analyses and of this paper is entirely on plant data sets. I suggest that the following are reasonable and *necessary* criteria in order to include a data set in a meta-study of the SRPR for plants.

1) Data must be provided for plant species richness and must be complete and consistent within the source paper. (Other diversity metrics may be of interest to ecologists, but the response variable should be the same throughout, so papers reporting other alpha diversity indices should be placed in a separate analysis.)

2) Plot size (and sampling regime) must be held constant (I suggest within ±10%, but with very small plots within ±5%) to avoid sampling variation confounding the analysis.

3) An adequate measure or surrogate for productivity must be available, which does not hold the danger of distortion of the relationship at high or low values, where most cases of unimodality are detected.

4) The data distribution (and spatial structure of the sampling) should be consistent with the assumptions involved in the statistical tests employed (and—although this isn't a criterion of inclusion/exclusion of data—those tests should in turn be appropriate and robust).

5) The study design should not involve significant variation internally in potentially confounding variables of known or strongly suggested importance, and which have a strong likelihood of invalidating the analysis (e.g., including differential impacts of mowing, grazing, horticulture, or burning that are correlated with the "productivity" gradient).

6) As data sets consisting of a very small number of data points can be insufficient to capture the form of the SRPR reliably, a minimum qualifying number of plots should be set at the outset. Given that the goal is to discriminate linear from unimodal form, Gillman and Wright (2006) adopted a 10 data point minimum, which seems a reasonable (but admittedly arbitrary) minimum for present purposes, and which I endorse.

7) The same data points should not be included either wholly or in substance more than once. This may seem an obvious criterion, but it is one that needs careful checking given the habit in ecology of reanalyzing data sets to different ends in different papers.

(One reviewer commented about criterion 3 that it is surely necessary to standardize productivity measurements in regard to the division between aboveground and belowground productivity. This is a fair point. Most studies report only a measure of aboveground productivity, and while little is known about belowground productivity in many systems, there is reason to suspect that across some ecological clines, there can be strongly differential patterns of allocation switching between above- and belowground biomass [C. Girardin, *personal communication*]. While I have noted this point, I have not added it to the numbered list of criteria as I have not attempted to apply it herein.)

I recognize that these criteria are hard to meet, and that few studies are available that meet them (Mittelbach et al. 2003), but so be it. If the data aren't appropriate to meta-analysis, it is invalid to proceed with one. The solution is to read the literature, think about it, and do one of the following: (1) devise some critical experimental or other rigorous field study that will make a meaningful contribution to the question to hand, (2) undertake a narrative review, or (3) carry out what Slavin (1995) has termed "best evidence synthesis." For a description of what this final technique embodies, see Slavin (1995).

Of the three meta-analyses under consideration, Gillman and Wright (2006) have the most stringent and explicit criteria (with common elements to the seven I have listed), while Pärtel et al. (2007) have the least explicit and most liberal approach to inclusion of data sets. Unfortunately, all three meta-analyses include data sets that should have been excluded, in the case of Gillman and Wright (2006) and Pärtel et al. (2007) this even extends to accidentally including the same data set twice.

## A Critical Audit of the SRPR Based on 68 Plant Data Sets from the Meta-Analyses

### *The method*

I selected 68 studies previously classified in one or more of the meta-analyses for critical re-evaluation. First, I made use of the limited re-analysis and reclassification provided by Whittaker and Heegaard (2003; designated WH2003) of data sets for trees classed by Mittelbach et al. (2001) [designated M2001] as "regional" or "continental-global" in scale ($n = 12$). Second, I selected papers haphazardly from the Appendix of Pärtel et al. (2007; designated P2007), as this was the most recent of the meta-analyses. Initially, I focused on those SRPR classed by P2007 as unimodal (humped), as this was where the greatest problems were detected by Gillman and Wright (2006; designated GW2006). I extended the selection in order to ensure a reasonable representation of different SRPR forms as defined by P2007, subject to ease of retrieval of the article pdf. I continued collecting papers until I reached 68 studies,

Table 2. Summary of comparisons of the classification of the form of the species richness–productivity relationship (SRPR) across 68 data sets in the three published meta-analyses and in this paper.

| Studies A vs. B | In study A only | In study B only | In A and B | Same result | Similar result | Different result | Percentage different |
|---|---|---|---|---|---|---|---|
| M vs. GW | 0 | 5 | 35 | 5 | 5 | 25 | 71.4 |
| M vs. P | 6 | 30 | 30 | 18 | 0 | 12 | 40.0 |
| M vs. RJW | 0 | 30 | 34 | 5 | 1 | 28 | 82.4 |
| GW vs. P | 4 | 25 | 36 | 7 | 2 | 27 | 75.0 |
| GW vs. RJW | 0 | 23 | 39 | 29 | 3 | 7 | 18.0 |
| P vs. RJW | 0 | 8 | 57 | 10 | 4 | 43 | 75.4 |

*Notes:* Values in each column represent the number of studies, with the exception of the final column, which represents the percentage of those classified in both papers that are put into different classes. Note that each pair-wise comparison involves differing subsets of data sets and thus different total *n* values (column labeled "In A and B"). Key to studies: M, Mittelbach et al. (2001), consensus classification; GW, Gillman and Wright (2006); P, Pärtel et al. (2007); RJW, R. J. Whittaker (this paper). "Same result" indicates same classification in both papers; "Similar result" indicates broadly the same outcome, but e.g., described as uncertain, or being a relationship with biomass rather than productivity; "Differing result" indicates a differing classification; "Percentage different" is the percentage of those classified in both papers that are put into different classes.

which although fewer than compiled by M2001 (121 plant data sets), GW2006 (159), and P2007 (163), is sufficient to establish the consistency and reliability of the meta-analyses.

My approach to the re-evaluation took the form of two not entirely separable elements, first the application of the above criteria, and second, an evaluation based on the analyses and contextual information presented in the original source paper of the form of the SRPR. My method did not involve any statistical reanalysis, but took the form of scrutiny of the aims, methods, sampling strategy, results, and discussion of the original papers to determine the validity of the classification applied in each of the meta-analyses. I contend that as long as the evidential basis of this process is made transparent and explicit, this form of scrutiny of the internal consistency and ecological logic of each original analysis provides powerful evidence on which judgments can be taken. Accordingly, I provide as my evidence key details of the properties of each data set and how the SRPR was classified in the meta-analyses (see Appendix A). Of course, this is in essence just a first step. Note that many of the source papers either did not attempt to test the form of the SRPR or failed to carry out analyses that compared unimodal and linear models in a directly comparable fashion. To improve the power of my audit, it would be necessary to test model fits directly on the original data, using the stipulation adopted by M2001 that unimodal fits should only be accepted when the maximum (humped) or minimum (U-shaped) value in the fitted quadratic term falls within the empirical range of the observed data. There are other important issues (e.g., is the quadratic term in fact significant?) and both WH2003 and GW2006 have shown that the statistical procedure adopted for this evaluation is important. However, they have also shown that greater problems have arisen from failings of basic experimental design criteria, inappropriate treatment of surrogate productivity variables (which are often not fit for the purpose) and of ecological logic. If we can't be sure of the meaning and validity of the data being entered for statistical analysis, disputation over statistical protocols merely distracts attention from the really serious problems. In practice, it turns out that only a few of the data sets are fit for purpose (Appendix A).

### *The emergent outcome*

Table 2 summarizes the outcome of the classification of the 68 data sets in the three meta-analyses and in my own audit. I should stress that it is tricky working out in some cases what particular data set is being referred to within P2007 and to a lesser degree in GW2006. Some data sets have been attributed to different source papers by different meta-analyses and in cases the same analysis has been included twice (below, Appendix A). Additionally, a few recent studies have provided analyses of the same system at multiple focal scales (e.g., Chase and Leibold 2002, Braschler et al. 2004, Chalcraft et al. 2004) and P2007 have been inconsistent in the number of "votes" assigned to these studies. Hence, as some papers provide multiple data sets, or multiple scales of analysis, and other data sets are included across the overall data base multiple times, the number of data sets could be deemed to be either more or less than 68.

The total number of data sets being directly compared varies between 30 (M2001 vs. P2007) and 57 (P2007 and RJW [this paper]) and the percentage of cases where the classification of the SRPR is different among the formal meta-analyses varies from 40% (M2001 vs. P2007) to 75% (GW2006 vs. P2007). Comparisons involving my own classification show that I largely concur with GW2006 (82% of cases), but I reject M2001's decisions in 82% of cases. In fact, in 17 of the 29 cases where GW2006 and RJW agree, we each classify the studies as inadmissible (i.e., invalid), meaning that we agree on only 12 meaningful classifications of SRPRs. From

Table 3. Summary of how analyses of the 68 data sets compare in their overall classification of studies, ignoring uncertain classifications, those deemed species richness–biomass relationships in Gillman and Wright (2006) and other such complexities (detailed in Appendix A).

| Paper | Positive | Humped | Negative | U-shaped | Inadmissible |
|---|---|---|---|---|---|
| Original paper | 8 | 12 | 3 | 0 | NA |
| Mittelbach et al. (2001) | 1 | 22 | 5 | 4 | 1 |
| Gillman and Wright (2006) | 6 | 5 | 0 | 2 | 21 |
| Pärtel et al. (2007) | 15 | 34 | 0 | 0 | 0 |
| RJW (this paper) | 5 | 7 | 3 | 0 | 35 |

*Notes:* Different subsets of the 68 data sets are included in each meta-analysis and my process of selecting studies may not have resulted in a representative subset of each meta-analysis. As a result, this table provides only a crude illustration of the way in which different approaches taken by each set of authors may have shaped the outcome of their analyses. Shape forms are as defined in *Introduction*; "Inadmissible" means failing the criteria outlined in the *Introduction* (i.e., invalid). NA = not applicable.

within 32 studies/data sets common to all three, the three meta-analyses concur in the classification of just three cases, make a very similar classification (i.e., basically opting for the same shape) in a further two cases, and disagree about the classification of 27 cases. So, at best, they agree on five studies, a meager 15% of the decisions. Of those five that have more-or-less agreed outcomes between the three meta-analyses, I dispute the classification of one more, meaning that across the four sets of authors, we have agreement on 4 studies (11%) out of 36 analyzed by each of us. You may as well classify the studies by random numbers. It is apparent that the meta-analyses of the SRPR provide no reproducible, objective basis for making any statement on emergent properties of the SRPR, how it varies with latitude (Pärtel et al. 2007), clonality of dominants (Laanisto et al. 2008), extent of study system (Mittelbach et al. 2001), and so on.

It is noteworthy that P2007 differ so much and analyze so many different data sets from M2001 because Pärtel et al. (2007) claim to have built their analyses largely on M2001 and because they provide no hint of how they classified the additional studies they included in their meta-analysis. Closer examination (Appendix A) suggests that their classification has also been influenced, to a limited degree, by GW2006 (and even by Whittaker and Heeagaard 2003). Part of the explanation for the difference in the classification of shared studies between P2007 and M2001 is that P2007 collapsed the initial five possibilities into three groups: (1) humped (including negative) SRPR, (2) positive SRPR, (3) no relationship (including U-shaped SRPR), of which more below. This resulted in four negative SRPR (M2001) being reclassified as humped SRPR (P2007). However, this collapsing of categories has not been carried out consistently. For instance, there are three cases where M2001 classified the SRPR as humped while P2007 didn't, and three of M2001's humps (each refuted by WH2003) were simply discarded from P2007's analysis. In addition, two of M2001's U-shaped relationships, instead of being reclassified by P2007 as "no relationship" were reclassified to a hump and a positive SRPR, respectively.

To begin to give some illustration of the breadth of the problems, taking the seven stated criteria listed above, in the Pärtel et al. (2007) paper: requirement 1 is broken in, e.g., case studies 9, 10, 11, 40, 129, 134, 135, 136, 137, 144, 145; requirement 2 in, e.g., cases 43, 66, 120, 129, 144; requirement 3 in, e.g., cases 8, 40, 46, 47, 62, 118, 120, 133; requirement 5 in, e.g., studies 8, 40, 51, 62, 91, 118, 133, 144, 146, 157; requirement 6 in, e.g., cases 62, 84, 91; requirement 7 in cases 106 and 108 (the exact same data set), and across the meta-analyses in other cases. Requirement 4 is also broken but is not formally demonstrated in my (nonstatistical) analysis other than by comparison across meta-analyses (see, e.g., study 147 in Appendix A). Appendix A demonstrates that there are more cases I could add to each list, those selected simply being "nice" examples. Unfortunately, very many of the above issues and examples apply also to the M2001 analysis, from which P2007 derived a large number of their classifications (Appendix A).

My sampling of the data sets used in the meta-analyses was not random in any formal sense but in a majority of cases I had not previously read the papers I selected for my reanalysis and so did not know in selecting them what I would find. In the earlier critique by Whittaker and Heegaard (2003) we provided a refutation of eight supposedly humped SRPR claimed by M2001, leading to the counter-charge that we were engaged in special pleading against humps (Mittelbach et al. 2003). So, I would like to emphasize that in this critique I do not simply dispute humped SRPR in this paper, and agree that they do occur (Table 3). However, it is apparent from close reading of the source material that both M2001 and P2007 are far too generous toward the notion of humped SRPR and far too liberal in assigning SRPR form without proper basis (Table 3, Appendix A). From the high level of erroneous and inconsistent treatments (both between and within meta-analyses) encountered for the 68 data sets examined, I anticipate that auditing of the remaining cases in the meta-analyses would reveal many additional errors and invalid classifications.

### *A few tasters*

Limitations of space mean that I can provide just a few potted examples in the main text, as follows. Ehrman and Cocks (1990) provide data concerning the distribution of annual legumes in Syria, organized as a form of percentage incidence based on varying numbers of sites from 12 climate zones. The paper thus provides no proper species richness data, is focused only on a small taxonomic subset, lacks standardized sampling across the gradient, and includes in the study design confounding variables, but is classified using rainfall variation into humped SRPR by M2001 and P2007. It is clearly inadmissible. O'Brien's (1993) data set for southern African trees is included twice in both GW2006 and P2007. It is also wrongly classed as a humped SRPR by M2001, as shown by WH2003 and supported by GW2006 and P2007 (twice in each case!). Across the three meta-analyses the same data set is sourced to three original papers, and two different sets of meta-data are provided for this single richness vs. rainfall data set. Flenley's (1969) phytosociological study of the vegetation of the Wabag Hills (New Guinea highlands), including gardens, swamps, disturbed forest, and undisturbed forest is included (uniquely) in P2007 despite the obviously unsuitable nature of the "experimental design" of this study system and the absence of a meaningful productivity surrogate and inadequate size of the forest plots involved. Wardle et al. (1997) is classed as a humped SRPR by P2007, despite the fact that the *islands* concerned varied in area across two orders of magnitude, the source paper lacks species richness data and has no productivity data (stand biomass was used as surrogate). Wheeler and Shaw (1991) report a negative SRPR explaining 36% of the variation in a data set for herbaceous rich-fen vegetation from the United Kingdom. M2001 and P2007 regard it as a humped relationship, while GW2006 classify it as U-shaped, and claim incorrectly that M2001 did the same. But these descriptions really are just tasters. Please see the fuller accounts of all 68 studies in Appendix A, read the source papers, and judge for yourself.

### *Humps by proclamation?*

Many of the studies included by Pärtel et al. (2007) were not conceived of as anything to do with the SRPR, and as several were not included in prior meta-analyses it is a mystery how the data were extracted, manipulated, analyzed and contextualized. As previously mentioned, however, Pärtel et al. are explicit that they started with five groups of relationships, which they then collapsed to three groups. First, "… the negative productivity–diversity relationship was merged with the unimodal relationship because most studies reporting a negative correlation focused on intermediate and high productivities" (Pärtel et al. 2007:1093). That is, they assume that all negative SRPR are merely incomplete humps in which the initial upward limb was (by design or accident) not sampled by the original authors. This is an extraordinary thing to do (1) as it invokes a complete reversal in the trend found in a data set by reference to no data at all, and (2) because the premise regarding productivity range is questionable (see, e.g., Appendix A, study 152: Wheeler and Shaw 1991). The merging of U-shaped relationships into the "no relationship" group on the grounds that U-shaped SRPR are theoretically implausible is also hard to justify (Gillman and Wright 2006), given that this is a mathematically equivalent form to their favored hump-shaped SRPRs. In fact, while they claim (Pärtel et al. 2007:1093) to have "used the earlier local and regional plant data from Mittelbach et al. (2001), but included additional studies (Appendix)," as shown above (and Appendix A) they have not simply incorporated the consensus decisions from M2001 but appear to have been influenced by decisions made both by M2001 and GW2006, while agreeing completely with neither. They have also, of course, added in further data sets scavenged from other source papers. Remarkably, I can find no trace in the paper or their appendix of the criteria and methods used to classify any of the SRPR in their meta-analysis. In addition, in at least two studies examined, it appears that an alpha diversity index was used instead of richness (Appendix A).

I fully accept that my own attempts to designate humps, U-shapes, and linear relationships in this article were based merely on visual examination of the source data and a reading of the source papers and whatever analyses they provide, but unlike M2001 and P2007, I explain the basis of my interpretation, I am explicit that the resulting designations are in cases highly uncertain, and I stress that they are not fit for summing for the purpose of further statistical analysis.

A final important point in attributing meaning to the SRPR is that having established, for instance, that a humped relationship is significant and explains more variation than a linear fit, if the overall variance explained is nonetheless very low, such that the majority of the variance in the data remains unexplained, this would suggests that something other than productivity is driving the system. Then the danger is that the apparent SRPR may in effect be an artifact of one or more other controlling factor(s): see Appendix A for discussion of a number of such cases.

## Why and How Focal Scale and Extent Are Important Organizing Principles

It appears from the body of literature reviewed herein (i.e., the source papers as well as the recent meta-analyses) that the understanding of scale and its significance to the analysis of phenomena such as the SRPR is very uneven and incomplete among the ecological community. There are three relevant components: grain, focus, and extent (Whittaker et al. 2001, 2003). (1) The grain refers to the basic sampling unit (e.g., plot) used in collecting the data, which must be appropriate to the task. (2) The focal scale refers to the inference space used in analysis, either simply being to make use of the basic sampling unit (in which case focal scale and grain are identical), or it may refer to a coarser scale to which the basic data are aggregated prior to analysis. (3) The extent refers to the geographical area within which the entire data set is bounded. Grain and focal scale are true scale components, whereas extent is not: increasing extent is equivalent to unfolding a map sheet, gradually revealing more of the region at a consistent resolution.

Regardless of the underlying grain of the original data, it is the unit used in analysis (i.e., the focal scale) that must be the primary organizing principle when it comes to comparisons across (between) studies. This is because, first, the larger the space enclosed in a sample, the more individuals and the more species is it liable to contain. In relatively species-poor systems it is possible for the species accumulation curve (the "sampling curve") to level fairly quickly, indicating that a local community has been adequately sampled. But, with further expansion in plot area to incorporate differing habitat type(s) (beta diversity) or species pools from different source regions (~gamma diversity), species numbers rise again, producing either stepped or smoothly rising curves depending on the heterogeneity of the study system (e.g., see Cody 1975). Particularly if the study unit size (grain) corresponds with steep phases of the species accumulation curve, it is crucial to hold the sampling unit exactly constant in order to avoid sampling effects confounding the analysis, and consideration should be given to aggregating sets of nearby sites together into a consistent, but coarser focal scale to minimize the likelihood of noise or of systematic bias entering the analysis. Failure to hold focal scale constant within a particular data set fatally compromises analyses using species richness, perhaps the most scale-dependent of ecological response variables (Whittaker et al. 2001, 2003, Rahbek 2005).

As previously commented by Whittaker and Heegaard (2003) a key weakness of the meta-analytical design used by Mittelbach et al. (2001) was that having undertaken their initial classification of each SRPR, they then organized their analysis by grouping studies into extent classes, instead of by focal scale: an approach they subsequently defended. This is to mix up entirely dissimilar sets of relationships and, I argue, entirely confounds their analysis. This misconception of the scale problem is widespread in the SRPR literature. For instance, Schamp et al. (2003) implicitly accept this prioritization of extent over grain in their paper, describing their own study as a regional scale study. However, while the extent of the system is truly regional (spanning several hundred km across southern Ontario), the grain size and focal scale used in the analysis is 10 × 10 m plots. These are small plots for forest communities, which at best may capture the local diversity, or alpha diversity (sensu Whittaker 1977) of the stand. One consequence of the grouping of data sets by extent rather than grain, is that Mittelbach et al. (2001) and other authors following this rationale, are trying to find pattern across sites spanning several orders of magnitude of (focal) spatial scale. It is highly likely that the most general property of the SRPR is that its form will be found to change as the grain/focal scale of the study system is changed (Whittaker et al. 2001, Chase and Leibold 2002, Whittaker and Heegaard 2003), especially when dealing with small plots, as a difference between one square meter and a few tens of square meters will often be crucial to the form of the relationship while changing resolution from 10 000 km$^2$ to 25 000 km$^2$ may turn out to have trivial impact (cf. Gillman and Wright 2006).

Holding focal scale constant in analysis is also desirable because each data point in a SRPR is stable in both the dependent and independent variable. Imagine that we have 20 study sites each of 1 m$^2$ scattered across an area of 1 km$^2$. If the extent of the study system is increased to 10 km$^2$ to capture a greater range in environment, the original 20 data points will be afforced by additional data points but their productivity and richness values are unaltered. Altering extent while holding focal scale constant thus allows us to "fill in" the statistical distribution, and if we have indeed captured a greater range in environment, we may well add data points disproportionately at one or both "ends" of the distribution (i.e., very high or very low productivity), aiding in the discrimination of (and perhaps changing) the form of the SRPR but not altering in any way the values and structure of our initial 20 data points. Imagine instead that within our large study system extent of 10 km$^2$ we increase the size of each sample plot, beginning always with the same central location point, from 1 m$^2$ to 4 m$^2$ to 20 m$^2$ and so on, and what might happen? The richness of each plot either increases or remains constant with each increase in plot size (decrease being impossible given that each larger plot contains the previous smaller one), while the productivity value assigned to the site can increase, remain the same, or decrease. This is because, unlike richness, which is an additive variable in this context, values of productivity may be averaged across a site, and can be lower on average in a 20-m$^2$ area than within a particular 1-m$^2$ patch within that 20-m$^2$ space. In general, we should expect a reduction in range of values of productivity as we increase the focal scale of our 20 data points, providing of course that we do use a true average for estimating productivity and do not, for instance, simply rely upon the same clipped sample of aboveground biomass in one particular place within each site.

The instability of values of independent and dependent variables means that the form of the SRPR can change rapidly and profoundly with shifts in focal scale of analysis, particularly where starting with very small plots. The corollary of this is that where researchers have set out to study the SRPR and providing a sensible sampling strategy has been adopted, using a fixed-size analytical unit (focal scale) within a given study area (extent), I would predict that a robust and relatively stable form of SRPR can quite quickly be established, so that adding additional plots makes little impact on the relationship. Changes to the form of that relationship can be anticipated, however, if either the study system is expanded in extent to encompass higher or lower productivity areas outside the geographic bounds of the original data set, or if the sampling protocol is altered so that distinct and different habitat types are added to the data set within the same system (geographical) extent (cf. Nogués-Bravo et al. 2008).

The logical conclusions of this line of argument are that first, in order to establish how the SRPR changes with variation in the *range* of climate, or productivity, or between biogeographical regions and so forth, it is system sampling strategy and/or geographical extent that should be altered while focal scale must be held constant, and second, that whatever pattern is established in the analysis holds true only for the focal scale used in that analysis and cannot be generalized to different focal scales. Recent studies that have used data for the same system extents, but aggregated to different focal scales, have shown that this second conclusion, which is derivable from first principles of ecological science, is also empirically true: the form of the SRPR varies with focal scale (Chase and Leibold 2002, Braschler et al. 2004, Chalcraft et al. 2004). This finding opens a further challenge. My requirement 7 (data points should only be included once in the meta-analysis) is designed to avoid bias introduced by double-counting the same system. But, what should be done with data for the same system that have been analyzed at different focal scales? How should they be treated in a meta-analysis? If, for instance, the focal scale is changed trivially, and the form of the SRPR is humped at the two adjacent scales, should both counts be included? That would seem to constitute double counting. But what if a third much coarser focal scale is provided, and now the result is a positive SRPR. How should this system be entered into the meta-analysis? Should one scale or one form of SRPR have precedence over the other? And, if so, on what rationale? Those undertaking meta-analyses cannot simply ignore this question if they wish to claim objectivity and repeatability for their analysis. For just such a case, see studies 16 and 17 (Braschler et al. 2004), in Appendix A.

There is one further point to be made concerning system extent. While increasing the geographical extent of the study system can increase the range of productivity values within an analysis, when comparing across different studies we should expect no simple relationship between extent and the range of values. For instance, in the lower middle latitudes, orographic features can produce pronounced variation in climatic conditions (water regimes and temperature, but not day length), and thus in productivity, in the space of a few kilometers, as can rivers running through the world's more arid areas. On the other hand, some data sets used in meta-analyses of the SRPR include tropical rain forest sites sampled in different continents, spanning a vast geographical extent but only a limited range of climate space. While such a data set does contain huge variation in terms of the constituent species pools involved in the different regions, the range of variation in the independent variable, i.e., productivity, may be quite limited compared with the local dry–mesic scenario outlined above. Grouping studies for analysis of the SRPR by their geographical extent that comprise data sets varying in their focal scale across many orders of magnitude, as undertaken by Mittelbach et al. (2001), is to generate an analysis fundamentally confounded by (true) scale. The empirical analyses by Chase and Leibold (2002) and Braschler et al. (2004) show this to be so.

Pärtel et al. (2007), on the other hand, simply ignore focal scale and system extent altogether, which is even worse, as *both* parameters are fundamental to the emergent form of the SRPR. Again, examination of two case studies is instructive. Chase and Leibold (2002) analyzed the richness of aquatic macrophytes in 30 ponds of about 500 $m^2$. They report a unimodal SRPR at the pond scale, but a simple positive SRPR when the data were aggregated up to the catchment scale by combining approximately three ponds per catchment. As Pärtel et al. (2007) were interested in analyzing variation in form of the SRPR with latitude, and failed to structure the analysis by scale, this particular study system appears twice in their meta-analysis for the same geographical co-ordinates, once as a unimodal SRPR (study 25, pond scale) and once as a positive SRPR (study 26, catchment scale), i.e., two different votes for the same place. In a second case study, Braschler et al. (2004) report analyses at three spatial scales, in each case providing separate analyses for graminoids, forbs, and forbs with graminoids. If following Braschler et al. (2004: Fig. 2) we could score this study as providing two unimodal relationships, four negative relationships and three null relationships. Or, we could follow the rationale that including taxonomic subsets of the same data is a form of "double-dipping" and we could just include the combined data for forbs with graminoids ("all plants") at each of three reported scales, providing one unimodal, one null and one negative relationship. In this case, P2007 enter two unimodal records for this study system (their studies 16 and 17), i.e., two rather than three "votes" for this system. A third example of multi-scale analysis is the paper by Chalcraft et al. (2004), who provide two focal scales of analysis for two separate sites, providing potentially four "votes": recognized by GW2006, but not by P2007 who record two "votes" only for this system. Hence, multi-scale treatments have been handled in different ways within P2007 and across the different meta-analyses. In fact, while the Braschler et al. (2004) study has other important things to say, the key conclusion to emerge from each of these three source papers is that there is no single form of SRPR for the systems they have analyzed and crucially the outcome is dependent on the focal scale used in the study.

P2007 not only make no attempt to control scale effects, they do not even record the scale parameters of the study systems in their meta-analysis. Their approach is to contrast the form of SRPR between low and high latitudes. But, as we know that the form of the SRPR varies depending on the focal scale used in analysis of the same data sets, and as focal scale (and extent) varies across many orders of magnitude in the studies compiled in each of the meta-analyses, it is nonsensical to undertake such an analysis. So, even were their initial classifications of the form of each particular study system correct (which in the great majority of cases they are not), their meta-analysis would be fatally compromised by the variation in the distribution of focal scale and ecosystem-scale properties between low and high latitudes in their study.

I just referred to "ecosystem-scale properties," by which I had in mind another largely intractable problem in analyses of the SRPR (Marañón and García 1997, Gillman and Wright 2006). How should we handle systems in which there is a mix of extremes of vegetation types, e.g., low-herbaceous grasslands and woodland? There are several such studies in the meta-analyses (e.g., Weiher 2003). Trees have a modular unit size that is orders of magnitude larger than grass and forb ramets. To move across a gradient from open areas to oak woodland, as in Weiher's (2003) study, is to traverse a gradient in which the effective physical and resource space available to herbaceous species becomes vastly reduced (cf. Oksanen 1996). Overall system net primary productivity (NPP) is likely to be highest in the tree-dominated stands, so how should we treat such study systems? Should we record all plant diversity and all NPP, or should we restrict our measurements of both to the herbaceous layer? If we do the latter, how should we account for the reduction in physical space and especially resources in the woodland quadrats? Effectively, the incursion of trees into the stands means that sampling/resource space for herbaceous species has not been held constant even though plot dimensions have been (for an extreme example see Nilsson and Wilson 1991, who used 0.5 × 1.0 m quadrats despite the fact that their system included 5 m high stands of *Betula pubescens*; Appendix A, studies 103, 104). If we include the trees in both measurements, on the other hand, we have crossed an important boundary in ecosystem properties and seen a shift in the relative proportion of biomass contributed by many small plants (in treeless plots) in favor of a very few large plants: is this system going to provide a meaningful representation of the relationship between species richness and productivity? I regard this question as posing an unanswered theoretical challenge. For the record, Weiher's (2003) approach was to focus just on the herbaceous layer, but as his statistical analyses showed, the SRPR was in any case compromised by the active fire regime of the study system. Partël et al. (2007) classify it as a humped SRPR.

## Plot Size Dictates the Form of the SRPR: A Theoretical Exposition

The question of what constitutes an acceptable minimum plot size is one that may depend in part on the purpose of the analysis, but it is surely self-evident that if your plot is too small to contain a single dominant individual, then it is too small to represent the local community (Gillman and Wright 2006). In a recent re-examination of appropriate plot sizes for phytosociological study of European vegetation, Milan and Zdenka (2003:563) come to the following conclusion: "... Based on our analysis, we suggest four plot sizes as possible standards. They are 4 $m^2$ for sampling aquatic vegetation and low-grown herbaceous vegetation, 16 $m^2$ for most grassland, heathland and other herbaceous or low-scrub vegetation types, 50 $m^2$ for scrub, and 200 $m^2$ for woodlands." Similar guideline plot sizes have in fact been around for decades, based largely on the wisdom that if the species accumulation curve for a vegetation type is beginning to approach an asymptote then a more-or-less stable representation of the local community may have been attained. Often, of course, plots need to be considerably larger than these sizes for stabilization of values to be reached (T. Stohlgren, *personal communication*). It is noteworthy that many of the studies used by Mittelbach et al. (2001) and by Pärtel et al. (2007) have plots significantly smaller than the least of these sizes (i.e., $<4$ $m^2$), including a number that were initially designed to analyze SRPR or species biomass–productivity relationships.

But, does this really matter? If the unit plot size is fixed, even if it is at a point on the species accumulation curve where richness is climbing steeply with increasing plot size, surely comparisons can be made? Yes, they can, but we should recognize that in such a case we are essentially working with point diversity (within community) rather than alpha diversity (richness representative of the local community) (sensu Whittaker 1977). This distinction may be important for interpretation of the SRPR (Oksanen 1996). Species accumulation curves typically rise very rapidly initially, and then flatten increasingly slowly until reaching an asymptote, rising again only when habitat boundaries are crossed to bring in genuine beta or gamma diversity (sensu Whittaker 1977) into the curve.

At very small plot sizes, beginning with perhaps 25-$cm^2$ grassland plots, physical competition for space, light, water, and nutrients is key in determining presence and richness of sub-patches within a sward. Using tiny plot sizes, we may therefore predict that analyses should typically return negative SRPR, as any increase in productivity will tend to be accompanied by a switch to larger ramets or clonal systems of one or two species (increased dominance, reduced equitability), reducing the likelihood of fitting in representatives of other species in these very small units of analysis. If there is an initial rising limb before the negative phase kicks in it will be apparent only briefly, with the negative phase starting at quite low productivity values. If we increase the plot size to and beyond the recommended Milan and Zdenka (2003) standards (i.e., a size where species accumulation curves are flattening) we can expect to see much more evidence of an initial rising limb as increased system productivity across a set of plots of varying productivity is matched by fitting in more ramets while retaining high equitability. Nonetheless, with further increases in productivity, we can again expect to find eventual decreases in species richness, particularly if our system includes artificially (or naturally) fertilized ("polluted") sites, within which those relatively few species in the local species pool that are best adjusted to exploiting high levels of nutrient are able to competitively out-grow other community members, expressing dominance (reducing equitability) and generating a reduction in richness. Thus, as we increase the focal scale, the position of the peak in richness should typically move from low in the productivity range toward higher values of productivity. And, as we escape the plot size at which local communities are defined and move to larger grain sizes (and different data types), and focal scales of analysis (up to and beyond 1000 km$^2$), we should expect to see increasing proportions of cases where species richness increases positively with productivity, either in a linear relationship or as an asymptotic curve with no downwards limb (Whittaker et al. 2001, Whittaker and Heegaaard 2003). This expectation is consistent with the overall findings of Gillman and Wright's (2006) meta-analysis, which I regard as the most rigorous of those reviewed herein.

To sum up, this theoretical exposition is linked to different conceptual realizations of diversity, invoking within-patch dynamics recorded at *point* scales of analysis, moving up to plot sizes more fully representing the local communities (i.e., to *alpha* scales of analysis), and eventually jumping to *gamma* scales of analyses, including whole landscapes or regions and in which climatic controls on species pools become apparent (in each case, point, alpha, and gamma are sensu Whittaker 1977). Hence, I posit that a lot of the variation reported in the literature on the form of the SRPR, in so far as it is based on adequate productivity data, and is meaningfully and accurately reported, essentially arises as an artifact or by-product of variation in the effective scale of sampling from point to alpha to gamma diversity. This argument is similar to but extends arguments made by Oksanen (1996). Variation in form at fine focal scales—and indeed what constitutes an appropriate scale of alpha analysis—will also depend on the range of physiognomic vegetation types incorporated (Marañón and García 1997, Chalcraft et al. 2004). As a crude generalization, however, negative SRPR should be expected to be most frequent for point scale data, with humped relationships more apparent at coarser alpha scales, and a gradual right shift of the hump, giving way to positive relationships within gamma scale analyses (shown schematically in Appendix B: Fig. B1). This somewhat speculative prediction could be tested by analyses using nested sampling based on plots of increasing grain size but fixed location across a fixed system extent.

## Concluding Comments

In my former role as editor-in-chief of *Global Ecology and Biogeography*, it was my idea to introduce Meta-Analysis as an article type in that journal: then I was fired up with enthusiasm for the approach, now I wonder if we should not have labeled the section "Here be dragons," as might be found on some ancient maps to describe unknown and generally hazardous regions of the world from which even the bravest explorers have rarely returned unscathed.

On my first theme—failings of prior analyses—I conclude that much of the original research undertaken within what has become a paradigmatic framing of humped SRPRs has been poorly designed experimentally, has involved strongly confounding variables, inadequate plot sizes, and poor choices of incomplete surrogate variables. Several of these themes, notably the highly problematic nature of biomass as a productivity surrogate (Gillman and Wright 2006, Keeling and Phillips 2007), have scarcely been touched on in this critique, while others are detailed only in Appendix A. I hope, however, to have demonstrated that enough of these problems are important, to demand a reappraisal of thinking on the SRPR. The meta-analytical contribution to understanding the SRPR started with a transparent but flawed analysis (Mittelbach et al. 2001), which, however, succeeded in knocking down the notion that the SRPR has a general form (and that this general form is humped), progressed with a worthy (but imperfect) reanalysis (Gillman and Wright 2006), and has proceeded to the point where there no longer seems to be any stated or reproducible criteria or method involved (Pärtel et al. 2007, Laanisto et al. 2008). Despite efforts to correct failings in the original meta-analysis (Whittaker and Heegaard 2003, Wright and Gillman 2006), further meta-analysis papers have appeared that mutate outcomes from Mittelbach et al. (2001), compound many of the original failings, and add new ones, a process of multiplying small errors to the point of producing wholly unsound outcomes. All sorts of entirely inappropriate data sets have now been recycled to answer questions that are incompletely specified and essentially unanswerable. Meta-analysis has led, in short, to mega-mistakes.

I have written this article not with any desire to fall out with those whose work I have criticized but because I happen to think an understanding of the SRPR is of considerable importance within ecological and biogeographical theory and because I feel that ecology as a discipline would be ill served by letting these chronic failings multiply through the literature unchecked. These failings in the treatment of scale, sampling design, plot size, and so on, in fact extend well beyond the meta-analyses, but at least these weaknesses are readily detectable in the original case study papers. Colleagues, we have to do better than this when we undertake and review and read and cite meta-analyses. Perhaps we can, in time, address the lack of standardization of experimental design and the tendency to change our methods from one study to the next, and find more reliable ways of dealing with the inherent multivariate nature of ecological systems, but in the meantime, we should be wary of trying to crunch (analyze) chalk and cheese data sets together, and we should be circumspect in regard to the use of meta-analysis in ecology.

On the second theme of this article, my case is that analyses of the SRPR that are not placed in an explicit scale framework are essentially meaningless. And, while the geographical extent of the system can influence form of the SRPR, it is intrinsically less problematic to compare studies of different system extent than to attempt to meta-analyze systems of differing focal ("true") scale of data: in fact to do the latter is nonsensical in the same way that it would be nonsensical to compare the diversity of a 1-m$^2$ patch of grassland to a 1-km$^2$ area of grassland. We know this from first principles and we now know it from empirical proof of the relevance of focal scale to the form of the SRPR.

I think an understanding of the variation in form of the SRPR must involve an understanding of the different processes at work at different scales of analysis and of how these are likely to structure our data sets. At fine scales of analysis we need to combine sampling theory with an understanding of species abundance distributions and species accumulation curves (see, e.g., Oksanen 1996, Marañón and García 1997, Chalcraft et al. 2004), and at all scales we have to deal with the multivariate nature of ecological processes. My proposition in this article is speculative, and incomplete theoretically, focusing as it does on largely artifactual mechanisms, but for what it is worth, predicts a general switch in form from negative and unimodal to positive SRPR with increasing focal scale of analysis. While collecting together and "crunching" (i.e., analyzing) large collections of data sets has its place in ecology (I am not entirely averse to it myself), we may advance faster in our understanding of this particular relationship by framing innovative primary studies designed to test particular hypotheses than by paying attention to the misleadingly precise quantifications generated by the meta-analyses.

## Acknowledgments

For discussion of these issues and/or comments on a presubmission draft I thank Bradford Hawkins, Kostas Triantis, Richard Ladle, Ole Vetaas, and David Nogués-Bravo; and for comments in review I thank Thomas Stohlgren, Gary Mittelbach (who kindly commented despite obvious differences of view), and two anonymous reviewers.

## Literature Cited

Beadle, N. C. W. 1966. Soil phosphate and its role in molding segments of the Australian flora and vegetation, with special reference to xeromorphy and sclerophylly. Ecology 47:992–1007.

Braschler, B., S. Zschokke, C. Dolt, G. H. Thommen, P. Oggier, and B. Baur. 2004. Grain-dependent relationships between plant productivity and invertebrate species richness and biomass in calcareous grasslands. Basic and Applied Ecology 5:15–24.

Chalcraft, D. R., J. W. Williams, M. D. Smith, and M. R. Willig. 2004. Scale dependence in the species-richness–productivity relationship: the role of species turnover. Ecology 85:2701–2708.

Chase, J. M., and M. A. Leibold. 2002. Spatial scale dictates the productivity–biodiversity relationship. Nature 416:427–430.

Cody, M. L. 1975. Towards a theory of continental species diversities: bird distributions over Mediterranean habitat gradients. Pages 214–257 *in* M. L. Cody and J. M. Diamond, editors. Ecology and evolution of communities. Harvard University Press, Cambridge, Massachusetts, USA.

Ehrman, T., and P. S. Cocks. 1990. Ecogeography of annual legumes in Syria: distribution patterns. Journal of Applied Ecology 27:578–591.

Flenley, J. R. 1969. The vegetation of the Wabag region, New Guinea highlands: a numerical study. Journal of Ecology 57: 465–490.

Gillman, L. N., and S. D. Wright. 2006. The influence of productivity on the species richness of plants: a critical assessment. Ecology 87:1234–1243.

Keeling, H. C., and O. L. Phillips. 2007. The global relationship between forest productivity and biomass. Global Ecology and Biogeography 16:618–631.

Laanisto, L., P. Urbas, and M. Pärtel. 2008. Why does the unimodal species richness–productivity relationship not apply to woody species: a lack of clonality or a legacy of tropical evolutionary history? Global Ecology and Biogeography 17:320–326.

Marañón, T., and L. V. García. 1997. The relationship between diversity and productivity in plant communities: facts and artefacts. Journal of Ecology 85:95–96.

Milan, C., and O. Zdenka. 2003. Plot sizes used for phytosociological sampling of European vegetation. Journal of Vegetation Science 14:563–570.

Mittelbach, G. G., S. M. Scheiner, and C. F. Steiner. 2003. What is the observed relationship between species richness and productivity? Reply. Ecology 84:3390–3395.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381–2396.

Nilsson, C., and S. D. Wilson. 1991. Convergence in plant community structure along disparate gradients: are lakeshores inverted mountainsides? American Naturalist 137: 774–790.

Nogués-Bravo, D., M. B. Araújo, T. Romdal, and C. Rahbek. 2008. Scale effects and human impact on the elevational species richness gradients. Nature 453:216–219.

O'Brien, E. M. 1993. Climatic gradients in woody plant species richness: towards an explanation based on an analysis of Southern Africa's woody flora. Journal of Biogeography 20: 181–198.

O'Brien, E. M. 1998. Water–energy dynamics, climate, and prediction of woody plant species richness: an interim general model. Journal of Biogeography 25:379–398.

O'Brien, E. M., R. J. Whittaker, and R. Field. 1998. Climate and woody plant diversity in southern Africa: relationships at species, genus and family levels. Ecography 21:495–509.

Odum, E. P. 1969. The strategy of ecosystem development. Science 164:262–270.

Oksanen, J. 1996. Is the humped relationship between species richness and biomass an artefact due to plot size? Journal of Ecology 84:293–295.

Pärtel, M., L. Laanisto, and M. Zobel. 2007. Contrasting plant productivity–diversity relationships across latitude: the role of evolutionary history. Ecology 88:1091–1097.

Pianka, E. R. 1966. Latitudinal gradients in species diversity: a review of concepts. American Naturalist 100:33–46.

Rahbek, C. 2005. The role of spatial scale and the perception of large-scale species-richness patterns. Ecology Letters 8:224–239.

Schamp, B. S., L. W. Aarssen, and H. Lee. 2003. Local plant species richness increases with regional habitat commonness across a gradient of forest productivity. Folia Geobotanica Phytotaxonomica 38:273–280.

Slavin, R. E. 1995. Best evidence synthesis: an intelligent alternative to meta-analysis. Journal of Clinical Epidemiology 48:9–18.

Wardle, D. A., O. Zackrisson, G. Hörnberg, and C. Gallet. 1997. The influence of island area on ecosystem properties. Science 277:1296–1299.

Weiher, E. 2003. Species richness along multiple gradients: testing a general multivariate model in oak savannas. Oikos 101:311–316.

Wheeler, B. D., and K. E. Giller. 1982. Species richness of herbaceous fen vegetation in Broadland, Norfolk, in relation to the quantity of above-ground plant material. Journal of Ecology 70:179–200.

Wheeler, B. D., and S. C. Shaw. 1991. Above-ground crop mass and species richness of the principal types of herbaceous rich-fen vegetation of lowland England and Wales. Journal of Ecology 79:285–301.

Whittaker, R. H. 1977. Evolution of species diversity in land communities. Pages 250–268 *in* M. K. Hecht, W. C. Steere, and B. Wallace, editors. Evolutionary biology. Volume 10. Plenum Press, New York, New York, USA.

Whittaker, R. J., and E. Heegaard. 2003. What is the observed relationship between species richness and productivity? Comment. Ecology 84:3384–3390.

Whittaker, R. J., K. J. Willis, and R. Field. 2001. Scale and species richness: towards a general, hierarchical theory of species diversity. Journal of Biogeography 28:453–470.

Whittaker, R. J., K. J. Willis, and R. Field. 2003. Climatic–energetic explanations of diversity: a macroscopic perspective. Pages 107–129 *in* T. M. Blackburn and K. J. Gaston, editors. Macroecology: concepts and consequences. British Ecological Society Symposia Series. Blackwell Publishing, Oxford, UK.

Williams, R. J., G. A. Duff, D. M. J. S. Bowman, and G. D. Cook. 1996. Variation in the composition and structure of tropical savannas as a function of rainfall and soil texture along a large-scale climatic gradient in the Northern Territory, Australia. Journal of Biogeography 23:747–756.

## APPENDIX A

Case-by-case evaluation of plant data sets used in three meta-analyses of the species richness–productivity relationship (*Ecological Archives* E091-185-A1).

## APPENDIX B

Schematic diagram of how changing focal scale may influence the form of the species richness–productivity relationship (*Ecological Archives* E091-185-A2).

FORUM

# Evidence and inference: shapes of species richness–productivity curves[1]

The relationship between productivity and species richness is a deeply embedded concept in ecology. There is little dispute that these two variables are usually positively correlated, at least from low to intermediate values. Species differ in resource use, more species can use a greater spectrum of the available resources, and this leads to greater productivity. However, from intermediate to high productivity and diversity, complicated powerful interactions of competition, consumption, disturbance, and spatial scales come into play. Either productivity or species richness, or one feeding back onto the other, can drive the relationship. Much of the new science—the endeavor of adding to our knowledge of nature—about the species richness–productivity relationship, SRPR, has played out in the pages of *Ecology* (see S. I. Dodson, S. E. Arnott, and K. L. Cottingham. 2000. The relationship in lake communities between primary productivity and species richness. Ecology 81:2662–2679 and B. J. Cardinale, D. M. Bennett, C. E. Nelson, and K. Gross. 2009. Does productivity drive diversity or vice versa? A test of the multivariate productivity–diversity hypothesis in streams. Ecology 90:1227–1241).

An influential meta-analysis (G. G. Mittelbach, C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381–2396) found that hump-shaped relationships between productivity and diversity were most common, while all forms, positive, negative, and no relationship, were represented in the literature. Perhaps the most important and frequently cited results of this analysis were two. First was the lack of uniformity in the relationship; humps were not "ubiquitous." Second was the importance of spatial scale; hump shapes were most frequent in large-scale studies that encompassed multiple communities and that included the greatest ranges of productivity. The study was not unchallenged, however, especially on the topic of spatial scale and quality of the data. In this Forum, we return to these fraught facets of SRPR with eight authors who broadly discuss the nature and validity of evidence and the means of inferring generality among studies in this relationship.

R. J. Whittaker leads with elaboration upon previously published critiques of the evidence and "calling time on meta-analyses" of SRPR. At the heart of Whittaker's points are recommendations of stringent data quality criteria. He argues that the lion's share of the problems with reviews of SRPR are in careless, indiscriminate inclusion of studies. Most of the responders share three messages. First is general agreement with the spirit, if not the particular details, of Whittaker's call for more attention to the data. The responders, especially Ellison, cite the need for data transparency, data depositories, universal data availability, and standards that will ensure quality of data and consistency of use (we are working toward these goals at ESA journals; see policy *available online*).[2] Second, most of the responders have great confidence in meta-analysis as the most coherent and scientifically powerful way to summarize and assess the commonalities and differences among studies of ecological phenomena in general and SRPR in particular. Finally, none of the responders sees a need to call time on meta-analysis in any area of ecology.

Aaron Ellison clears the air by pointing out the different activities in the efforts to reveal commonalities in the form of the SRPR curve: assembly of derived data sets, repeated reanalysis but not meta-analysis of these data, and finally, meta-analysis itself. While Mittelbach et al. (2001) conducted formal meta-analysis, the other studies targeted by Whittaker did not. Gillman and Wright counter Whittaker by arguing that studies of SRPR have not fallen "into chaos," and rather than abandoning meta-analyses, ecology should exercise greater caution in use of data. Mittelbach speaks softly and carries the big stick of the many citations that ecologists have given to Mittelbach

[1] Reprints of this 50-page Forum (including the focal Concepts and Synthesis paper by Whittaker) are available for $10.00 each, either as PDF files or as hard copy. Prepayment is required. Order reprints from the Ecological Society of America, Attention: Reprint Department, 1990 M Street, N.W., Suite 700, Washington, D.C. 20036 (esaHQ@esa.org).

[2] ⟨http://esapubs.org/esapubs/conditions.htm#Arc⟩

et al. (2001). He, like several other responders, counters that Whittaker's advocacy of narrative, "best evidence synthesis" is not a solution to data quality problems. His final point is also shared widely: meta-analysis is the best means to assess the validity of suspected problems in data, such as methodology, scale, study size, organism type, habitat, and others. It is the most scientifically powerful and ecologically insightful way to study commonalities and differences in SRPR among studies.

Hillebrand and Cardinale are skilled users of meta-analysis and know its strengths and weaknesses from long experience. They state that authors of meta-analyses are obliged to appreciate natural history and to "read each paper included in their analyses carefully and to understand the unique features of a study that might influence one's conclusions." They make the important point that meta-analysis in ecology is no stranger to data challenges, and that these techniques have a good track record of improvement through time. The power of meta-analyses is in "moving beyond patterns," and scientific customization of meta-analyses will reveal ecological mechanisms and specific predictions of ecological theory. As other respondents, Hillebrand and Cardinale are highly critical of retreat into case-specific, narrative analysis. At the same time, one feels their suspicion of studies that rest on sheer volume of data, have unclear hypotheses, lack statistical rigor, and are vague about mechanisms.

Gurevitch and Mengersen explain how responses nearly identical to Whittaker's have been seen in other disciplines, as in medicine for example. They reject, point by point, unwarranted assertions about meta-analysis and explain the failings of less quantitatively coherent techniques, such as "vote counting." Lajeunesse explains the two-edged sword of stringent data standards. Power and generality decrease with the elimination of studies, and one can erroneously eliminate studies from an analysis just as one can erroneously include them. He argues that instead of striving for data purity by pruning studies from a data set, why not include all studies and empirically weigh the contribution of each of the perceived deficiencies to the outcome, such as to the quadratic parameter for SRPR curves estimated among studies by meta-analysis? Finally, with lengthy appendices, Whittaker details specific complaints about data, and Pärtel et al. defend their work against Whittaker's criticisms with lengthy appendices of their own.

This exchange will be a rich guide to the ways ecologists do science, and it will provide some hints for both experienced researchers and those just entering the field about how not to proceed. We should now move on to the vital task of accomplishing universal data and metadata deposition, with user friendly software, and unfettered access, detailed by Ellison in his response to Whittaker.

—Donald R. Strong
*University of California–Davis*

*Key words: data quality criteria; hump-shaped relationship; meta-analysis; productivity; species richness; SRPR.*

*Ecology*, 91(9), 2010, pp. 2536–2539
© 2010 by the Ecological Society of America

# Repeatability and transparency in ecological research

Aaron M. Ellison[1]

*Harvard University, Harvard Forest, 324 North Main Street, Petersham, Massachusetts 01366 USA*

## Introduction

A fundamental tenet of science is that results must be reproducible by other scientists before they are accepted as factual. However, because ecological phenomena are context-dependent, and because that context changes through time and space, it is virtually impossible to reproduce precisely or quantitatively any single experimental or observational field study in ecology. Yet many ecological studies can be repeated. In particular, *ecological synthesis*—the assembly of derived data sets and their subsequent analysis, reanalysis, and meta-analysis—should be easy to repeat and reproduce. Such syntheses also demonstrate qualitative and quantitative consistency among many ecological studies (Gurevitch et al. 1992, Warwick and Clarke 1993, Jonsen et al. 2003, Walker et al. 2006, Cardinale et al. 2006, Marczak et al. 2007, Vander Zanden and Fetzer 2007) and provide strong support for general ecological theories.

It should come as no surprise that meta-analysis by Mittelbach et al. (2001) of the effect of productivity on species richness has led to the development of a cottage industry focused on empirical testing of this relationship (post-2001 examples abound in Appendix A of Whittaker 2010). But it is much more surprising that continual reanalyses of the *same* data sets (Whittaker and Heegaard 2003, Gillman and Wright 2006, Pärtel et al. 2007) have yielded such disparate results that Whittaker (2010) has suggested abandoning the effort to obtain consistent results from the available data. He goes even further, suggesting that ecology may not yet be ready for meta-analysis and data synthesis. For two reasons, I respectfully suggest that Whittaker's critique is misplaced. First, of all the studies critiqued by Whittaker (2010), only Mittelbach et al. (2001) actually conducted a formal meta-analysis. The others, as pointed out by Whittaker (2010), undertook extensive primary analyses, but the authors did not conduct formal meta-analyses (Gurevitch and Hedges 1999). Second, and more importantly, if ecological synthesis is transparent—data, models, and analytical tools are available freely to the research community—then it should yield consistent, repeatable results. We may then disagree on the *interpretation* of the resulting synthesis, but at least we will be able to agree on the reproducibility of the results themselves.

## Requirements for Repeatable Ecological Synthesis

In a nutshell, ecological synthesis proceeds by assembling available data sets into a common, derived data set and then applying one or more (statistical) models to this derived data set to test the prediction of a hypothesis of interest (Ellison et al. 2006). Repeatability and reproducibility of ecological synthesis requires full disclosure not only of hypotheses and predictions, but also of the raw data, methods used to produce derived data sets, choices made as to which data or data sets were included in, and which were excluded from, the derived data sets, and tools and techniques used to analyze the derived data sets. Of all the papers under discussion by Whittaker (2010), Mittelbach et al.'s (2001) paper comes closest to achieving such transparency, although neither the raw data nor the derived data set they analyzed are publicly available.

But achieving this level of disclosure and transparency is difficult. First and foremost, researchers must be committed to transparent production of ecological knowledge. We may be blissfully unaware of our own intellectual biases, but there are no excuses for not making data, methods, and tools freely available in a timely fashion. Yet despite mandates from funding agencies and research networks that data be made available publicly (Arzberger et al. 2004), raw data are not easily accessed. Research teams can spend many weeks searching data archives only to find summary statistical tables, lists of means, or concise graphs. Contacting individual investigators may yield raw data in digital form or in yellowing notebooks, or it may yield nothing at all. Fortunately, archives of ecological data are growing (examples include ESA's data registry,[2] *Ecological Archives*,[3] the data repository of the National Center for Ecological Analysis and Synthesis [NCEAS],[4] the data archive of the Long-

Manuscript received 8 January 2009; accepted 17 June 2009. Corresponding Editor: D. R. Strong. For reprints of this Forum, see footnote 1, p. 2534.

[1] E-mail: aellison@fas.harvard.edu

[2] ⟨http://data.esa.org/esa/style/skins/esa/index.jsp⟩

[3] ⟨http://www.esapubs.org/archive/⟩

[4] ⟨http://knb.ecoinformatics.org/knb/style/skins/nceas/⟩

Term Ecological Research Network,[5] and Oak Ridge's Distributed Active Archive Center,[6] among many others), but archiving ecological data is not yet a requirement for publication in any journal. Ecologists also have developed standard methods for describing ecological data sets with *descriptive metadata* (Michener et al. 1997, Jones et al. 2006, Madin et al. 2008) that make it easier to interpret and hence re-use them. Software tools such as Morpho that help investigators create descriptive metadata also are maturing (software *available online*).[7]

But it is not enough simply to find a data set and understand its origin and structure. Once data sets are obtained, it is usually necessary to transform the data into common units and scales (e.g., species/ha or kg/ha). Interpolated values may need to be substituted for missing data, and methods of interpolation will vary among investigators (Ellison et al. 2006). Finally, and usually after still further manipulations and making decisions as to which data to include or exclude (cf. Whittaker and Heegard 2003, Whittaker 2010: Appendix A), a derived data set is ready for analysis.

Each step—e.g., digitization, rescaling, interpolation, inclusion, or exclusion—requires individual judgment and provides an opportunity to introduce bias or error. If subsequent synthesis is to be repeatable, users must have confidence in the reliability of the derived data set. Thus it is imperative that researchers document clearly each of the steps used to produce derived data sets. This *process metadata*—the documentation of the processes used to produce a data set—provides one way to assess the reliability of a derived data set (Osterweil et al. 2005, Ellison et al. 2006). Storage of the original data sets *and* the processes applied to create the derived data set provides the mechanism to reproduce it.

Such audit trails that include archived data sets and tools allow can allow future users to determine effects of changing particular processes on the structure and subsequent analysis of the derived data set (Ellison et al. 2006). For example, Mittelbach et al. (2001) classified the relationship between species richness and productivity in one of five categories (unimodal humped or U-shaped, monotonic positive or negative, or no relationship) whereas Laanisto et al. (2008) classified this same relationship simply as unimodal or not. Whittaker and Heegard (2003) and Whittaker (2010) excluded data that Mittelbach et al. (2001) included. Gillman and Wright (2006) used some of the regression results reported by Mittelbach et al. (2001) but also reanalyzed some of the original data sets using different software and without specifying which data were reanalyzed. Clearly results will differ if the same data are classified differently, if different subsets of data are analyzed, or if individual data sets are treated differently. Importantly, we can assess these differences by running new analyses on available data sets. The resulting differences in approach to and analysis of the data may reflect differences in questions on the part of the investigators, honest disagreements regarding the "best" available evidence (sensu Slavin 1995), or strongly held opinions regarding the most appropriate statistical analysis (e.g., ordinary least-squares regression vs. general linear models with a variety of error distributions and link functions). However, these differences and disagreements do not in and of themselves invalidate the activity of ecological synthesis.

It is equally important to document and whenever possible archive the statistical tools and models used for analysis and synthesis (Thornton et al. 2005); such an archival record should be a requirement for publication of any meta-analysis or data synthesis. The various authors critiqued by Whittaker (2010) all used different statistical tools (Table 1), and it would be impossible to repeat precisely any of the author's analyses.

Documentation and archiving of analytical processes, including those processes used to create derived data sets and the statistical tools and models applied to them, is difficult, and software tools for such documentation and archiving are rudimentary. It may seem wasteful to archive software, but numerical precision of arithmetic operations changes with new integrated circuit chips and different operating systems, functions work differently in different versions of software, and implementation of even "standard" statistical routines differ among software packages (a widely unappreciated example of relevance to ecologists is the different sums of squares reported by SAS, S-Plus, and R for analysis of variance and other linear models; Venables 1998). Finally, there are no standards for process metadata (Osterweil et al. 2005, Ellison et al. 2006) and no easy way to archive model code used by, or specific versions of, commercial software packages. While open-source software tools such as R (R Development Core Team 2007) are attractive (and affordable) alternatives, they evolve even more rapidly than their commercial counterparts, and regular changes in functionality of familiar routines are not uncommon (implementation of the cor function for calculation of Pearson's correlation coefficient in early versions of R is a notorious example). But without archiving software, tools, and associated process metadata, it is unlikely that we will be able to accurately reproduce any ecological synthesis.

## Moving Forward

More and more ecologists are following federal guidelines (Office of Management and Budget 1999) and making their data freely available within a short time of collection and publication (for analysis and agency-specific implementation of this regulation, see assessment at The Center for Regulatory Effectiveness

[5] ⟨http://metacat.lternet.edu/knb/⟩

[6] ⟨http://daac.ornl.gov/⟩

[7] ⟨http://knb.ecoinformatics.org/morphoportal.jsp⟩

TABLE 1. Analytical methods used in the syntheses of the species richness–productivity relationship.

| Author | Analytical method(s) used | Analytical tool(s) used | Comments |
|---|---|---|---|
| Waide et al. (1999) | linear and quadratic regressions | none specified | not repeatable |
| Mittelbach et al. (2001) | ordinary least-squares regression | SYSTAT 8.0 | possibly repeatable; current available version is 12.0 |
| | Poisson regression | NAG statistical add-in for Excel | not repeatable; software discontinued |
| | "Mitchell-Olds and Shaw test" (Mitchell-Olds and Shaw 1987) | none specified | not repeatable; software unavailable (but algorithm available); which of three tests proposed by Mitchell-Olds and Shaw was also not specified |
| | chi-square exact test | StatXact | possibly repeatable; no version given |
| | meta-analysis using mixed-effects model | MetaWin 2.0 | repeatable; commercial software version still available |
| Whittaker and Heergard (2003) | Poisson regression | not specified | not repeatable |
| Gillman and Wright (2006) | ordinary least-squares regression on "some" data sets of Mittelbach et al. (2001) | software not specified; data sets reanalyzed not specified | not repeatable |
| Pärtel et al. (2007) | multinomial logit regression | Statistica 6.1 | possibly repeatable; current release is 8.0 |
| Laanisto et al. (2008) | Fisher exact tests | not specified | possibly repeatable using available algorithms |
| | general linear model | Statistica 6.1 | possibly repeatable; current release is 8.0 |

*Note:* Manufacturers of software are: SYSTAT 8.0, Systat Software, Inc., Chicago, Illinois, USA; NAG statistical add-in for Excel, Numerical Algorithms Group, Oxford, UK; StatXact, Cytel, Inc., Cambridge, Massachusetts, USA; MetaWin 2.0, Sunderland Associates, Inc., Sunderland, Massachusetts, USA; Statistica 6.1, StatSoft, Inc., Tulsa, Oklahoma, USA.

Web site, *available online*).[8] Cultural impediments to data sharing among ecologists are disappearing as more and more ecologists recognize not only that sharing of data benefits the entire scientific enterprise (Baldwin and Duke 2005) but also results in successful collaborations and subsequent publications such as those facilitated by NCEAS (*available online*).[9] Rapid development of data archiving and sharing tools has been facilitated by funding initiatives focused on development of software for production of descriptive metadata and distributed access to permanently and stably archived data (see National Science Foundation, Office of Cyberinfrastructure, *online*).[10] There is increasing recognition that similar efforts must be undertaken to document analytical tools and processes and to archive the software tools themselves (Thornton et al. 2005, Ellison et al. 2006). Software tools in development for creating process metadata, including documentation of data set provenance and storage of analytical tools applied to derived data sets, include Kepler (Ludäscher et al. 2006) and the Analytic Web (Osterweil et al. 2010). Ecologists should work with these software development teams, and others like them, to learn how better documentation and archiving of scientific processes and work flows can advance our science and to provide challenging tests of these evolving systems (Boose et al. 2007).

Rather than abandon data synthesis and meta-analysis as Whittaker (2010) suggests, ecologists should embrace these activities as the very essence of our science. With appropriate attention to documentation of data *and* analytical processes and a commitment to unbiased inquiry and full transparency of analytic activities, data synthesis, and meta-analysis will become the most repeatable and reproducible activities that ecologists undertake. The results of such syntheses and meta-analyses will be the grist for the mill of ecological forecasting, perhaps the most important endeavor of 21st century ecology (Clark et al. 2001).

## ACKNOWLEDGMENTS

Gary Mittelbach discussed availability of the original species richness–productivity data set and Tom Mitchell-Olds answered questions about the availability of his Pascal software written in 1987. Brad Cardinale provided helpful comments on early versions of the manuscript. Work on this manuscript was supported by the Analytic Web project (NSF grant CCR-0205575) and by the Harvard Forest Long-Term Ecological Research Program (NSF grant DEB 06-20443).

## LITERATURE CITED

Arzberger, P., P. Schroeder, A. Beaulieu, G. Bosker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir, and P. Wouters. 2004. An international framework to promote access to data. Science 303:1777–1778.

Baldwin, J. D., and C. Duke. 2005. Society summit on data sharing and archiving policies. Bulletin of the Ecological Society of America 86:61–66.

Boose, E., A. M. Ellison, L. J. Osterweil, R. Podorozhny, L. Clarke, A. Wise, J. L. Hadley, and D. R. Foster. 2007.

[8] ⟨http://thecre.com/access/index.html⟩

[9] ⟨http://nceas.ucsb.edu/products⟩

[10] ⟨http://www.nsf.gov/dir/index.jsp?org=OCI⟩

Ensuring reliable datasets for environmental models and forecasts. Ecological Informatics 2:237–247.

Cardinale, B. J., D. S. Srivastava, J. E. Duffy, J. P. Wright, A. L. Downing, M. Sankaran, and C. Jouseau. 2006. Effects of biodiversity on the functioning of trophic groups and ecosystems. Nature 443:989–992.

Clark, J. S., et al. 2001. Ecological forecasts: an emerging imperative. Science 293:657–660.

Ellison, A. M., L. J. Osterweil, J. L. Hadley, A. Wise, E. Boose, L. Clarke, D. R. Foster, A. Hanson, D. Jensen, P. Kuzeja, E. Riseman, and H. Schultz. 2006. Analytic webs support the synthesis of ecological data sets. Ecology 87:1345–1358.

Gillman, L. N., and S. D. Wright. 2006. The influence of productivity on the species richness of plants: a critical assessment. Ecology 87:1234–1243.

Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. Ecology 80:1142–1149.

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

Jones, M. B., M. P. Schildhauer, O. J. Reichman, and S. Bowers. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics 37:519–544.

Jonsen, I. D., R. A. Myers, and J. M. Flemming. 2003. Meta-analysis of animal movement using state-space models. Ecology 84:3055–3063.

Laanisto, L., P. Urbas, and M. Pärtel. 2008. Why does the unimodal species richness–productivity relationship not apply to a woody species: a lack of clonality or a legacy of tropical evolutionary history? Global Ecology and Biogeography 17:320–326.

Ludäscher, B., I. Altintas, C. Berkeley, D. G. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, and Y. Zhao. 2006. Scientific workflow management and the Kepler system. Concurrency and Computation: Practice and Experience 18: 1039–1065.

Madin, J. S., S. Bowers, M. P. Schildhauer, and M. B. Jones. 2008. Advancing ecological research with ontologies. Trends in Ecology and Evolution 23:159–168.

Marczak, L. B., R. M. Thompson, and J. S. Richardson. 2007. Meta-analysis: trophic level, habitat, and productivity shape the food web effects of resource subsidies. Ecology 88:140–148.

Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Nongeospatial metadata for the ecological sciences. Ecological Applications 7:330–342.

Mitchell-Olds, T., and R. G. Shaw. 1987. Regression analysis of natural selection: statistical influence and biological interpretation. Evolution 41:1149–1161.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381–2396.

Office of Management and Budget. 1999. Office of Management and Budget Circular A-110, revised 11/19/93, as further amended 9/30/99. Executive Office of the President of the United States of America. ⟨http://www.whitehouse.gov/omb/circulars/a110/a110.html⟩

Osterweil, L. J., L. A. Clarke, A. M. Ellison, E. Boose, R. Podorozhny, and A. Wise. 2010. Clear and precise specification of ecological data management processes and dataset provenance. IEEE Transactions on Automation Science and Engineering 7:189–195.

Osterweil, L. J., A. Wise, L. Clarke, A. M. Ellison, J. L. Hadley, E. Boose, and D. R. Foster. 2005. Process technology to facilitate the conduct of science. Pages 403–415 *in* M. Li, B. Boehm, and L. J. Osterweil, editors. Lecture notes in computer science: SPW 2005. Springer-Verlag, Berlin, Germany.

Pärtel, M., L. Laanisto, and M. Zobel. 2007. Contrasting plant productivity-diversity relationships across latitude: the role of evolutionary history. Ecology 88:1091–1097.

R Development Core Team. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Slavin, R. E. 1995. Best evidence synthesis: an intelligent alternative to meta-analysis. Journal of Clinical Epidemiology 48:9–18.

Thornton, P. E., R. B. Cook, B. H. Braswell, B. E. Law, W. M. Post, H. H. Shugart, B. T. Rhyne, and L. A. Hook. 2005. Archiving numerical models of biogeochemical dynamics. Eos 86:431.

Vander Zanden, M. J., and W. W. Fetzer. 2007. Global patterns of aquatic food chain length. Oikos 116:1378–1388.

Venables, W. N. 1998. Exegeses on linear models. Paper presented at S-Plus User's Conference, Washington, D.C., 8–9 October 1998. ⟨http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf⟩

Waide, R. B., M. R. Willig, C. F. Steiner, G. Mittelbach, L. Gough, S. I. Dodson, G. P. Juday, and R. Parmenter. 1999. The relationship between productivity and species richness. Annual Review of Ecology and Systematics 30:257–300.

Walker, M. D., et al. 2006. Plant community responses to experimental warming across the tundra biome. Proceedings of the National Academy of Sciences USA 103:1342–1346.

Warwick, R. M., and K. R. Clarke. 1993. Comparing the severity of disturbance: a meta-analysis of marine macrobenthic community data. Marine Ecology Progress Series 92: 221–231.

Whittaker, R. J. 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91:2522–2533.

Whittaker, R. J., and E. Heegaard. 2003. What is the observed relationships between species richness and productivity? Comment. Ecology 84:3384–3390.

*Ecology*, 91(9), 2010, pp. 2540–2544
© 2010 by the Ecological Society of America

# Understanding species richness–productivity relationships: the importance of meta-analyses

Gary G. Mittelbach[1]

*W.K. Kellogg Biological Station and Department of Zoology, Michigan State University, Hickory Corners, Michigan 49060 USA*

> *Hence our truth is the intersection of independent lies.*
>
> —R. Levins (1966)

In the above quote, Levins was referring to "truth" and "lies" in model building, however, I believe his comments are relevant to the analysis of empirical data as well. We all recognize that published papers differ in quality, even those that are predominantly descriptive. Whittaker (2010), in his critique of meta-analyses of species richness productivity relationship (SRPRs), argues that few of the studies used in past meta-analyses of SRPRs are fit for the purpose. This leads him to "call time" on any further meta-analyses of SRPRs and to denounce the findings of previous meta-analyses as unreliable. Whittaker (2010:2524) states, "If the data aren't appropriate to meta-analysis, it is invalid to proceed with one. The *solution* [my italics] is to read the literature, think about it, and do one of the following: (1) devise some critical experimental or other rigorous field study that will make a meaningful contribution to the question to hand, (2) undertake a narrative review, or (3) carry out what Slavin (1995) has termed 'best evidence synthesis.'" I would argue, however, that these alternatives do not provide a *solution* and that the past meta-analyses of SRPRs, despite their weaknesses and disagreements, have significantly advanced our understanding of these relationships. The literature on SRPRs is uneven in quality and heterogeneous in method: on that there is no doubt. But, it is what we have to work with. In the end, we make progress by scrutinizing our ideas in light of the available data. At the risk of sounding extreme, I suggest that even empiricists must look for "truth" at the intersection of independent "lies." Consider what we "knew" about SRPRs prior to the published meta-analyses.

### *Some history on the SRPR*

Thirty years ago, Grime (1979) noted that in the plant communities he studied, species richness first increased and then deceased as soil fertility and plant biomass increased (a "humpbacked" relationship). Rosenzweig and Abramsky (1993) similarly documented humped-shaped SRPRs for a variety of animal communities (and ecologists have puzzled over the cause of the descending limb of the "hump" ever since). In the 1990s, the consensus was that most SRPRs were hump-shaped (Rosenzweig and Abramsky 1993, Tilman and Pacala 1993, Huston 1994, Rosenzweig 1995, Grace 1999). For example, Tilman and Pacala (1993:23) stated that, "The available observational evidence (nine studies cited) supports the hypothesis that plant diversity is a unimodal function of productivity or of other measures of nutrient supply rates. We know of no cases in which (plant) diversity is a simple increasing function of productivity or nutrient supply." Huston and DeAngelis (1994:972) described the hump-shaped SRPR relationship as "ubiquitous" in plants, and Rosenzweig and Abramsky (1993:55) stated that "Within regions about the size of small to medium-sized nations, species diversity is often—perhaps usually?—a unimodal function of productivity (or some well-accepted index of it like rainfall or nutrient supply)." Interestingly, Wright et al. (1993) published a review of species–energy theory in the same book that includes the reviews by Rosenzweig and Abramsky (1993) and Tilman and Pacala (1993). Wright et al. (1993:67) stated that "The energy–richness relationship is clearly scale-dependent. On the global scale, richness increases monotonically with energy (four studies cited), whereas on small scales, richness is sometimes a peaked function of energy (five studies cited)." Thus, their conclusions presage those reached by subsequent meta-analyses, although they provided no data to support their assessment.

By the mid-1990s, there was a broad consensus that SRPRs were hump-shaped at most spatial scales and at least one prominent ecology textbook discussed the unimodal SPRR as the general pattern for plants (Begon et al. 1990:825). This prevailing view was based on a combination of theory and supporting examples drawn from the literature, although Abrams (1995) showed that positive and unimodal relationships could arise from the same theory, and as noted above, Wright et al. (1993) suggested that the observed SRPR varied with spatial extent. As part of an early NCEAS working group, we conducted a broad survey of the literature and subjected

Manuscript received 9 June 2009; accepted 29 July 2009. Corresponding Editor: D. R. Strong. For reprints of this Forum, see footnote 1, p. 2534.

[1] E-mail: mittelbach@kbs.msu.edu

the observed SRPRs to a standardized statistical analysis. This quantitative analysis showed that hump-shaped relationships, although common, were not "ubiquitous" (Mittelbach et al. 2001; see also a preliminary analysis in Waide et al. 1999). We concluded (p. 2385) that "Although our survey of the literature turned up many significant hump-shaped relationships, the proportion may be less than expected based on current ecological thought," and that "Our survey shows that both hump-shaped and positive productivity–species richness relationships are common in nature, and we suggest that perhaps too much attention has been focused on looking for 'humps'" (p. 2394). This was a fairly radical view at the time; certainly, some of our reviewers found it radical. Whittaker, in his critique, questions why Mittelbach et al. (2001) is often cited, given its weaknesses (in his opinion). I think Mittelbach et al. (2001) is cited because our meta-analysis helped shift the focus away from the "ubiquitous" unimodal curve and instead showed that there is considerable variation in the form of the SRPR, and that the predominant form of the SRPR may depend on spatial scale and the environment.

Whittaker believes that we were far too liberal in including studies in our 2001 review and that this casts serious doubt on our conclusions (and has lead to a compounding of errors in later studies). We were liberal in our inclusion of studies because we wanted to get away from the selected example approach of the times (e.g., see figures in Rosenzweig and Abramsky 1993, Tilman and Pacala 1993, Huston 1994). We did, however, explicitly define the criteria used to locate and select studies for our analysis, and we documented the statistical analyses and criteria used to judge the form of the SRPR (Mittelbach et al. 2001). Gillman and Wright (2006), in response to a suggestion by Whittaker and Heegaard (2003), subsequently reanalyzed the terrestrial plant data from our review, along with 37 additional studies. They selected "…only those studies that have used appropriate surrogates for productivity and adequate controls for confounding factors" (Gillman and Wright 2006:1237). Gillman and Wright (2006) used essentially the same criteria to judge the admissibility of studies as Whittaker advocates in his current critique and it is their shared criteria of admissibility that leads to their high level of "agreement" (92%; Whittaker 2010: Tables 2 and 3). Thus, Mittelbach et al. (2001) and Gillman and Wright (2006) provide two meta-analyses of the literature on SRPRs for terrestrial plants, one more selective than the other, and we can compare their conclusions.

Gillman and Wright (2006) found that humped and positive SRPRs occurred at similar frequencies at the local and landscape scale, and that at larger spatial extents (regional to continental), SRPRs were essentially always monotonically positive (my summary based on their Fig. 3). We (Mittelbach et al. 2001) concluded that, "In plants, hump-shaped relationships were especially common at smaller spatial scales" (p. 2391) and that "positive relationships were common for vascular plants at the largest spatial scale … and the odds of finding a positive relationship tended to increase with an increase in spatial extent" (p. 2392). We found more unimodal relationships than did Gillman and Wright (2006) for two reasons. First, we used general linear model (GLM) regression to analyze the source data, whereas Gillman and Wright used ordinary least squares (OLS) regression. The impact of using GLM vs. OLS regression on these data sets was noted in our original paper (Mittelbach et al. 2001: Table 2) and we provided the results of both GLM and OLS regressions in an online appendix (see discussion in Mittelbach et al. 2003, Whittaker and Heegard 2003, Gillman and Wright 2006). Second, many of the data sets that yielded humped relationships in our analyses were rejected by Gillman and Wright (2006), because they felt that the $x$-axis variable did not scale monotonically with productivity.

To me, the bottom-line from these two meta-analyses is that SRPRs at local and landscape scales tend to be hump-shaped or positive, and that at larger spatial extents, the relationship shifts to becoming positive (although often decelerating). Gillman and Wright (2006) see substantial differences between our analyses based on the percentage of studies falling into the different categories and they argue that humped SRPRs are over-represented in our study. Whittaker views both studies as flawed. However, I see progress in our understanding of the SRPRs, from the 1990s view that the hump-shaped relationship is "ubiquitous," to our current understanding that the form of the SRPRs relationship varies across spatial scales and across systems. Unlike Whittaker, I believe the meta-analyses published in *Ecology* have contributed to this understanding and that they have inspired significant new work. These meta-analyses are far from perfect, but they are not "mega-mistakes."

### *Comparing the meta-analyses*

Whittaker criticizes both the quality of the data used in the published meta-analyses and the fact that they are inconsistent in the way they classify the relationships, leading him to conclude that the results are not repeatable and that one could do as well by classifying studies at random. Whittaker reached this conclusion by examining 68 plant data sets extracted from the three meta-analyses. These data sets included 28%, 24%, and 35% of the studies in the meta-analyses of Mittelbach et al. (2001), Gillman and Wright (2006), and Pärtel et al. (2007), respectively. He does not consider the 89 animal data sets in Mittelbach et al. (2001). The 68 studies examined by Whittaker were not selected at random (see his selection criteria on page 2524), and it is worth nothing that they include a high proportion of studies classified as unimodal; i.e., the cases where Whittaker and Heegaard (2003) and Gillman and Wright (2006) had the most disagreements with the analysis of

Mittelbach et al. (2001). Whittaker didn't statistically analyze the form of the SRPR reported in any study, but instead his re-evaluation "...took the form of scrutiny of the aims, methods, sampling strategy, results, and discussion of the original papers to determine the validity of the classification applied in each of the meta-analyses" (p. 2525). Basically, Whittaker examined each study and decided whether it met his criteria for providing a valid SRPR and then decided what the form of that relationship was based on visual inspection of the data, coupled with his and the original author's interpretation.

Whittaker reports the results of this re-evaluation in Tables 2 and 3, and concludes that there is so little agreement among the three meta-analyses and with his own analysis, that no meaningful conclusions can be drawn from the meta-analyses: "It is apparent that the meta-analyses of the SRPR provide no reproducible, objective basis for making any statement on emergent properties of the SRPR..." (p. 2526). Whittaker reports (p. 2525) that "...I reject M2001's decisions in 82% of cases." However, it turns out that 62% of these rejections (21 of the 34 studies examined) are because Whittaker classifies as "inadmissible" a study that we included in our meta-analysis. Of the remaining 13 studies, we agree on the shape of the SRPR in four, find a "similar result" in one, agree on one as "inadmissible" (below minimum sample size), and disagree on the form of the SRPR in the remaining six. (I realize that one study is missing from this tally, but I'll be darned if I can determine which one is based on Table A1.) Of the six studies where we disagree on the form of the SRPR, we classified five of the studies as humped or U-shaped; Whittaker classified four of them as monotonic (+ve or −ve) and one as (?). As discussed earlier, our GLM analysis tended to classify a greater percentage of SRPRs as humped or U-shaped compared to OLS regression.

I apologize for leading you through this long accounting, but I wanted to show you where Whittaker's statement that "I reject M2001's decisions in 82% of cases" comes from. I was surprised when I first saw this level of disagreement, but I understand its origins now. Interestingly, Whittaker's own tally of his 15 admissible studies (out of the 68 studies selected for analysis), includes seven studies (47%) that he classified as hump shaped, five studies (33%) classified as positive, and three studies (20%) classified as negative (Whittaker 2010: Table 3). Should we conclude from Whittaker's analysis of the "acceptable" literature that hump-shaped SRPRs occur more often than positive SRPRs in plants? No, probably not. Most likely, this result simply reflects the non-random way in which the 68 studies were selected. However, the result shows that even the most skeptical approach to the data finds a strong representation of both hump-shaped and positive SRPRs in nature.

### *Alternative approaches*

How should we best synthesize studies that are heterogeneous in quality or differ in sampling regime or experimental design? This is a fundamental question that extends beyond the study of SRPRs and it lies at the crux of this forum. A discussion of this general topic is best left to the experts (see papers in this Forum). Here, I only want to comment on Whittaker's suggestion that when the published data on diversity relationships are not appropriate for a meta-analysis, the solution is instead to (1) do an experiment or rigorous field study, (2) write a narrative review, or (3) carry out a "best evidence synthesis" (Slavin 1995). Point 1: no single study, no matter how well-designed and executed, can answer the question of how diversity relationships vary between taxa, or how they vary across geographic regions or between different habitats (e.g., freshwater vs. terrestrial). These questions can only be addressed by comparing the results from many, independent studies. Experiments also are necessarily limited in size and temporal scale. Therefore, I assume that Whittaker is not suggesting that experiments or field studies can substitute for a meta-analysis of SRPRs, but rather that they can contribute to understanding SRPRs in other ways. Point 2: Whittaker does not define a "narrative review," but presumably it involves a synthesis of the literature without extensive statistical analysis. In terms of the summarizing what was known about SRPRs, the reviews by Rosenzweig and Abramsky (1993), Tilman and Pacala (1993), Huston (1994), Rosenzweig (1995), and especially Grace (1999) for herbaceous plant communities, are good examples of this type of narrative review. As was noted above, these narrative reviews reached similar conclusions about the ubiquitous nature of the hump-shaped SRPR, whereas subsequent meta-analyses found the form of the SRPRs to be more varied and to differ with spatial extent and to differ between habitats. Point 3: best evidence synthesis. Slavin (1995), writing as part of a forum on meta-analysis in the *Journal of Clinical Epidemiology*, describes an approach to reviewing the literature and drawing conclusions that he terms "best evidence synthesis." Slavin (1995) suggests (and I am crudely summarizing here), that when the literature on a topic is uneven in quality, the best approach is to throw out the bad studies and retain the good ones, and then conduct a formal meta-analysis (e.g., a comparison of effect sizes). The best evidence approach differs most from standard meta-analysis in being selective rather than inclusive in choosing the studies for review, and in presenting a lengthy discussion of the individual studies that merit inclusion. One important question, of course, is "what to leave in, what to leave out" (apologies to Bob Seger). Slavin (1995) readily acknowledges the potential danger of introducing bias into a review, both in determining which studies are "good" and "bad" methodologically, and in (unconsciously) selecting studies that support a particular point of view.

Whittaker presents the three approaches above as solutions to the problem of how to deal with the heterogeneous and uneven literature on SRPRs, arguing that they are preferred to any past (and future) meta-analyses of SRPRs. I see these approaches as useful alternatives for probing the relationship between species richness and productivity, but I do not see a simple solution and would argue instead that multiple approaches (including meta-analyses) are needed.

### *Focal scale and extent*

How and why the form of the SRPR varies with spatial scale is a second theme in Whittaker's critique. Whittaker's comments about the potential effects of plot size and focal scale on the form of the SRPR are valuable and they follow closely an earlier conceptual paper by Whittaker and colleagues that nicely shows how different forms of the SRPR may theoretically arise as a function of changing diversity and productivity at different spatial scales, and by varying patterns of species turnover across the landscape (Whittaker et al. 2001). We (Scheiner et al. 2000) published on some of the same ideas, asking how SRPRs may differ as a function of scale (grain, focus, and extent). In his current critique, Whittaker discusses again the potential effects of sampling scale on the SRPRs, and chastises Mittelbach et al. (2001) and Pärtel et al. (2007) for not being sufficiently rigorous in controlling for variation in plot size. He also criticizes Mittelbach et al. (2001) for looking at how the form of the SPRP's changes with spatial extent rather than grain size. These criticisms were first raised in Whittaker and Heegaard (2003), to which we provided a response (Mittelbach et al. 2003). Since then, Gillman and Wright (2006) specifically examined how the form of the SRPR for terrestrial plants varies as a function of grain size (fine and coarse grain). They found that positive SRPRs predominated at both grain sizes, and that "...fine-grain studies produced more unimodal and nonsignificant relationships than coarse-grain studies" (p. 1237).

The larger part of Whittaker's discussion of scale effects focuses on the general issue of how grain, focus, and extent, in combination with species turnover across the landscape, may affect the form of the SRPR (i.e., Whittaker et al. 2001: Fig. 1). I agree that these are important issues. I disagree, however, with his premise that if changing grain or focal size *can* affect the form of the SPRP, then any study that includes variation in these parameters, or any compilation of studies (meta-analysis) that includes variation in these parameters, is fatally compromised. It is one thing to argue that a factor may affect an analysis, and another to demonstrate that it does. For example, Scheiner et al. (2000) discuss how species–area curves can (in principle) be used to standardize sampling scale between studies and they illustrate theoretically how aggregating sampling scales along a productivity gradient may affect the form of the SRPR. They then analyzed two empirical data sets (one for old fields in Michigan and one for tallgrass prairie watersheds in Kansas) to address this issue. They found that across sampling scales from 10 $m^2$ to 200 $m^2$, the species–area curve was rank invariant amongst sites of differing productivities. Therefore, for these two data sets, there would be no change in the overall shape of the SRPR due to changing grain sizes across a productivity gradient. I am not suggesting that changing grain size or focal scale can not affect the form of the SRPR; there are clearly examples where it does (e.g., Chase and Leibold 2002). However, to say that we must not undertake a meta-analysis of SRPRs because of these issues, strikes me as extreme. Whittaker argues that our 2001 meta-analysis is flawed because it examines extent and does not control for grain size (although we did test for plot size effects and didn't find any), and he argues that Pärtel et al. (2007) is flawed because they didn't explicitly consider either grain or extent. But, Gillman and Wright (2006) did limit their analysis to studies similar in grain size and they conducted the grain-size comparison that Whittaker calls for. Therefore, the comparison is available in the literature and readers can determine if these scale effects are damning to meta-analysis or not.

Whittaker goes on to postulate that in general plot size dictates the form of the SRPR (p. 20) and "that variation in the form of the SRPR at fine scales of analysis owes much to artifacts of the sampling regime adopted" (p. 1). This is a reasonable hypothesis for terrestrial plants, and it corresponds directly to the prediction illustrated in Fig. 1f in Whittaker et al. (2001). This argument is similar to Oksanen's (1996) hypothesis that an increase in plant size with fertility explains the reduction in species richness found in small sampling plots (due to self-thinning; but see ensuing counter arguments, e.g., Grime 1997, Maranon and Garcia 1997, Zobel and Liira 1997). Thus, the argument that plot size may influence the form of the SRPRs is not new. Moreover, it is a hypothesis that needs to be tested, not an empirical fact that demonstrates a fatal flaw in meta-analyses of SRPRs. It is interesting that the study of Chase and Leibold (2002), which Whittaker holds out as a clear example of how increasing focal scale changes the form of the SRPR from humped to positive, is a study of plant (macrophyte) and invertebrate species in freshwater ponds. In this study, the data on plant species richness were collected from transects (not plots), and the animal data were collected from cumulative samples with sweep nets. Thus, Chase and Leibold's result showing a change in the shape of the SRPR with an increase in focal scale can not be an artifact of plot size. Moreover, Chase and Leibold (2002) show that the only way for the SRPR to change form from hump shaped to positive as focal scale increases from local to regional (they aggregated the data from individual ponds (three) within each watershed to get the "regional" pattern), is for the dissimilarity in species composition between ponds (beta diversity) to increase with productivity. Therefore, the fact that the SRPR changes with focal scale in this study is not simply due to increasing the size of the sampling unit, but is instead driven by some underlying mechanism (as yet unassigned) that causes beta diversity to increase with productivity.

### *Conclusions*

Whittaker's critique raises important issues and it has served as a stimulus to a broader discussion on the use of meta-analysis in ecology. Meta-analysis was still quite new to ecology when we conducted our review of SRPRs (see Introduction to *Ecology* Special Feature; Osenberg et al. 1999). In Mittelbach et al. (2001), we conducted only one formal meta-analysis, addressing whether the standardized quadratic regression coefficients for SRPRs (from OLS regression) were more heterogeneous than expected by chance: they were and the overall mean quadratic coefficient was significantly negative, indicated that the average SRPR was nonlinear and decelerating. However, for the most part, our analysis consisted of comparing the frequency of different forms of the SRPR between different systems (e.g., aquatic vs. terrestrial, plants vs. animals) and across different spatial extents or community boundaries. As such, our review was a meta-analysis only in the broadest sense of the word, as it did not include an analysis of effect size (something we felt was impossible given the heterogeneity in the source data and study methods).

Looking back, I recognize that our analysis of SRPRs was only a first step (and an imperfect one at that). Yet, I stand by the statement that our study helped move the field forward, as have subsequent meta-analyses of SRPRs. We know much more about SRPRs today than we did in the mid-1990s, and we are beginning to see a synthesis of the factors controlling species richness and productivity from both sides of the relationship (e.g., Cardinale et al. 2009). Progress in science is made by confronting ideas with data. Caution and skepticism are needed. But, the only way to determine whether "here be dragons" or not, is to take a look.

### Acknowledgments

I thank Laura Gough, Kay Gross, Sam Scheiner, Chris Steiner, and two anonymous reviewers for thoughtful comments on the manuscript. This is contribution #1534 from the W. K. Kellogg Biological Station.

### Literature Cited

Abrams, P. A. 1995. Monotonic or unimodal diversity–productivity gradients: what does competition theory predict? Ecology 76:2019–2027.

Begon, M., J. L. Harper, and C. R. Townsend. 1990. Ecology. Second edition. Blackwell Scientific. Boston, Massachusetts, USA.

Cardinale, B. J., D. M. Bennett, C. E. Nelson, and K. Gross. 2009. Does productivity drive diversity or vice versa? A test of the multivariate productivity–diversity hypothesis in streams. Ecology 90:1227–1241.

Chase, J. M., and M. A. Leibold. 2002. Spatial scales dictates the productivity–biodiversity relationship. Nature 416:427–430.

Gillman, L. N., and S. D. Wright. 2006. The influence of productivity on the species richness of plants: a critical assessment. Ecology 87:1234–1243.

Grace, J. B. 1999. The factors controlling species density in herbaceous plant communities: an assessment. Perspectives in Plant Ecology, Evolution and Systematics 2:1–28.

Grime, J. P. 1979. Plant strategies and vegetation processes. John Wiley and Sons, New York, New York, USA.

Grime, J. P. 1997. The humped-back model: a response to Oksanen. Journal of Ecology 85:97–98.

Huston, M. A. 1994. Biological diversity: the coexistence of species in changing landscapes. Cambridge University Press, Cambridge, UK.

Huston, M. A., and D. L. DeAngelis. 1994. Competition and coexistence: the effects of resource transport and supply rates. American Naturalist 144:954–977.

Levins, R. 1966. The strategy of model building in population biology. American Scientist 54:421–431.

Maranon, T., and L. V. Garcia. 1997. The relationship between diversity and productivity in plant communities: facts and artefacts. Journal of Ecology 85:95–96.

Mittelbach, G. G., S. M. Scheiner, and C. F. Steiner. 2003. What is the observed relationship between species richness and productivity? Reply. Ecology 84:3390–3395.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381–2396.

Oksanen, J. 1996. Is the humped relationship between species richness and biomass an artifact due to plot size? Journal of Ecology 84:293–295.

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

Pärtel, M., L. Laanisto, and M. Zobel. 2007. Contrasting plant productivity–diversity relationships across latitude: the role of evolutionary history. Ecology 88:1091–1097.

Rosenzweig, M. L. 1995. Species diversity in space and time. Cambridge University Press, Cambridge, UK.

Rosenzweig, M. L., and Z. Abramsky. 1993. How are diversity and productivity related? Pages 52–65 *in* R. E. Ricklefs and D. Schluter, editors. Species diversity in ecological communities. University of Chicago Press, Chicago, Illinois, USA.

Scheiner, S. M., S. B. Cox, M. Willig, G. G. Mittelbach, C. Osenberg, and M. Kaspari. 2000. Species richness, species–area curves and Simpson's paradox. Evolutionary Ecology Research 2:791–802.

Slavin, R. E. 1995. Best evidence synthesis: an intelligent alternative to meta-analysis. Journal of Clinical Epidemiology 48:9–18.

Tilman, D., and S. Pacala. 1993. The maintenance of species richness in plant communities. Pages 13–25 *in* R. E. Ricklefs and D. Schluter, editors. Species diversity in ecological communities. University of Chicago Press, Chicago, Illinois, USA.

Waide, R. B., M. R. Willig, C. F. Steiner, G. Mittelbach, L. Gough, S. I. Dodson, G. P. Juday, and R. Parmenter. 1999. The relationship between primary productivity and species richness. Annual Review of Ecology and Systematics 30:257–300.

Whittaker, R. J. 2010. Meta-analysis and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91:2522–2533

Whittaker, R. J., and E. Heegaard. 2003. What is the observed relationship between species richness and productivity? Comment. Ecology 84:3384–3390.

Whittaker, R. J., K. L. Willis, and R. Field. 2001. Scale and species richness: towards a general, hierarchical theory of species diversity. Journal of Biogeography 28:453–470.

Wright, D. H., D. J. Currie, and B. A. Mauer. 1993. Energy supply and patterns of species richness on local and regional scales. Pages 13–25 *in* R. E. Ricklefs and D. Schluter, editors. Species diversity in ecological communities. University of Chicago Press, Chicago, Illinois, USA.

Zobel, K., and J. Liira. 1997. A scale-independent approach to the richness vs biomass relationship in ground-layer plant communities. Oikos 80:325–332.

*Ecology*, 91(9), 2010, pp. 2545–2549
© 2010 by the Ecological Society of America

# A critique for meta-analyses and the productivity–diversity relationship

Helmut Hillebrand[1,3] and Bradley J. Cardinale[2]

[1] *Institute for Chemistry and Biology of the Marine Environment, Carl-von-Ossietzky University Oldenburg, Schleusenstrasse 1, 26382 Wilhelmshaven, Germany*
[2] *Department of Ecology, Evolution and Marine Biology, University of California–Santa Barbara, Santa Barbara, California 93106 USA*

It is an exciting time to be an ecologist. Over the past several decades, our discipline has matured from one focused on the assembly of case studies based on natural history, to one that has seen improved conceptual frameworks and mathematical models that help explain ecological phenomena from species coexistence to elemental cycling. The maturation of our discipline has been fostered by many things, including improved technology, increased availability of data, and emergent methods for analyzing large data sets. One factor that has played a central role in the modern synthesis is meta-analyses. Gurevitch et al. (1992) introduced meta-analyses to ecologists and catalyzed their entrance into the ecological literature as a powerful statistical means to assess the generality of pattern and process. Soon after, the U.S. National Science Foundation established the National Center for Ecological Analysis and Synthesis (NCEAS) whose mission it is to bring together ecological data sets so that we could synthesize pattern and process using meta-analysis and many other analytical tools. NCEAS was so successful that it was soon after mimicked by other scientific disciplines (e.g., NESCent, the National Evolutionary Synthesis Center).

However, when our initial honeymoon with "synthesis" was over, criticisms began to surface, exposing the inherent warts and flaws of a growing discipline. Some argued that data sets were now being analyzed, and syntheses performed, by researchers who knew little about (or perhaps had never even seen) the systems they were trying to understand. Such "remote ecology" reduces an appreciation for natural history, and may lead to incorrect conclusions because one doesn't understand the intricacies and contingencies of each system that reveal how pattern is linked to process. Some argued that meta-analyses were proliferating more rapidly than the methods needed for quality control, and that the concatenation of data sets was leading to a propagation of errors.

These are essentially the arguments levied by Whittaker (2010, from now on W2010). W2010 strongly criticizes the use of meta-analyses in ecology, and uses three syntheses of a fundamentally important ecological pattern (the productivity-diversity relationship, PDR) to illustrate why he believes there are flaws in the use of this tool. He especially criticizes the first of these analyses by Mittelbach et al. (2001) (from now on M2001). He argues that these meta-analyses have lacked stringent and transparent criteria in study selection, have ignored important correlates of the relevant independent and response variables (e.g., spatial scale), and have been inconsistent in their categorization of studies to the extent that the authors of the different syntheses have reached divergent conclusions.

We agree with certain elements of Whittaker's criticism, including the need for improved quality control and transparency in the selection and analysis of data. We also agree with W2010's general sentiment that those who are performing meta-analyses have an obligation to read each paper included in their analyses carefully and to understand the unique features of a study that might influence one's conclusions. There is no substitute for thoroughly understanding the natural history of the system(s) from which one is drawing inference, and no substitute for characterizing the unique and shared features of the studies included in a synthesis (for example, see Foster et al. 2006's response to Halpern et al.'s 2006 meta-analysis, or the comments on Worm et al. 2006, including those by Holker et al. 2007, Jaenike 2007, and Wilberg and Miller 2007). We detail comments on these points in the section *Quality issues in meta-analysis*.

However, beyond his relatively straight-forward call to improve the way we conduct meta-analyses, there is little in W2010 that we agree with. We are especially worried that W2010 proposes that we throw out the baby with the bathwater, calling for a halt in meta-analyses so that we can refocus attention on the intricacies of each case study. This proposition thoroughly neglects the many improvements in the handling and analyses of data that have been developed for meta-analyses over the last 15 years (see accompanying

Manuscript received 14 January 2009; revised 17 July 2009; accepted 28 July 2009. Corresponding Editor: D. R. Strong. For reprints of this Forum, see footnote 1, p. 2534.

[3] E-mail: hillebrand@icbm.de

comments by Ellison [2010] and Gurevitch and Mengersen [2010]). It also suggests that ecological patterns and processes are so highly system specific that it is difficult, perhaps impossible, to extract general trends amid the background of natural variation. We couldn't disagree more, and we are generally enthusiastic that ecology as a discipline has moved beyond the case studies and contingencies of individual systems to seek generality (see *A general critique for meta-analysis*).

In our final section, *Moving beyond patterns*, we add to W2010's commentary by suggesting that much of the confusion and disparity in conclusions among those seeking to synthesize the PDR stems from a lack of clear mechanistic thinking. Summarizing patterns without a clear mechanistic understanding of theoretically plausible relationships does nothing other than lead to confusion, no matter how rigorous and technically sound a meta-analysis might be. Therefore, we end with an appeal to those who might perform further meta-analyses to think more deeply about what should, according to ecological theory, be the dependent and independent variables behind the productivity–diversity relationship.

## Quality Issues in Meta-Analyses

Objective and clear criteria for data inclusion are the cornerstone of any endeavor to synthesize data. A meta-analysis has to be based on the most comprising and unbiased set of studies affiliated with the research question at hand. W2010 is justified in stating that those performing meta-analyses sometimes do not do a very good job in stating their search and inclusion criteria. Based on the "Methods" section in a meta-analysis, a researcher should be able to redo the entire analysis starting with the literature search and database build-up, proceeding with the statistical analysis, and finally coming to the same conclusions. Of course, these criteria are no different than the standards that should be imposed by reviewers on any publishable research.

W2010 details what he believes are flaws in the three analyses investigated by him, and goes on to propose seven "improved" criteria for the inclusion of studies in future meta-analyses. We do not take issue with his claim about flaws: for example, that authors have selected and categorized data in different ways that are sometimes not transparent and repeatable, and that authors have occasionally made mistakes or double entries into their data sets, which have led researchers to divergent conclusions. Moreover, subsequent papers criticized the original meta-analysis for flaws not only in the database, but also that the statistical models used to analyze the data were inappropriate. Such issues are clearly important to resolve. We agree that there needs to be improved standards for quality control in meta-analytic data sets. This is an issue that has been discussed at length (Osenberg et al. 1997, Englund et al. 1999, Gurevitch and Hedges 1999, LaJeunesse and Forbes 2003, Rosenberg 2005), and there are ongoing attempts to develop the cyber-infrastructure needed to improve the management, sharing, and analysis of data in the next century (for example, the International Knowledge Network for Biocomplexity and accompanying management software Morpho). We also believe that debate over the nature and validity of different analyses is a normal, healthy part of the scientific process, and that this debate gradually leads to an improvement in conclusions. As such, W2010's comment helps promote a worthwhile debate.

However, we think it would be a tragedy to adopt W2010's strict criteria for how to overcome these problems. It generally strikes us as dangerous and naïve when a researcher suggests there is a single best, or optimal way to gain knowledge. Rather than trying to force researchers into a narrow mold, we believe that the primary constraints on a meta-analysis should be (1) *clarity*, researchers need to clearly state how the data is being used and why; (2) *transparency*, researchers to make abundantly clear how the data were collected, which data were included, and why; (3) *technical accuracy*, researchers need to be sure that the assumptions of their statistical tools match the structure of the data; and (4) *availability*, researchers need to make all data and technical code from their analyses available along with the published results so that the accuracy, reliability, and repeatability of the data set can be checked.

Although we have no doubt that W2010's comment was written in an attempt to promote clarity, accuracy, transparency and availability, his proposed criteria overshoot this aim by restricting analyses to very narrow grounds at the expense of the larger picture potentially gained by meta-analyses. For example, in criterion 1 W2010 argues that plant species richness is the only reasonable response variable that should be used as a measure of diversity. Although we agree that richness is the focus of many theories about productivity–diversity relationships, and although we agree that researchers should take great care to measure variables that are mechanistically consistent with ecological theory, there is no reason to believe species richness is the sole aspect of diversity that should be related to productivity. Not only is species richness itself a proxy for how other aspects of biodiversity are packaged (e.g., genetic or evolutionary divergence), several of the mechanisms by which productivity is thought to influence richness occur through changes in species evenness (Hillebrand et al. 2007), spatial turnover of taxa ($\beta$ diversity; Chase and Leibold 2002, Gross and Cardinale 2007), or changes in functional group dominance (Declerck et al. 2007). Rather than argue that ecologists focus narrowly on just one response variable, it would be more constructive to emphasize that the focal variable should be motivated by the question, and that the question itself should be constructed so that there is a clear mechanistic underpinning or theoretical justification for expecting the response variable to change with productivity.

In criteria 2 and 5, W2010 argues that the studies included in an analysis should be completely homogeneous with respect to the scales at which they are performed, and with respect to the wide variety of potentially confounding variables that might influence species richness and productivity. Aside from the fact that this demand is incredibly unrealistic and would prevent us from summarizing more than a handful of studies at any one time, this argument ignores the fact that it is often a far more powerful approach in synthesis efforts to maximize variation among studies so that one can determine which co-varying factors actually "matter" in a way that they alter conclusions about the form of productivity-diversity relationships. A meta-analysis is especially useful if it reveals that a conclusion holds across a broad variety of empirical approaches or, alternatively, if it shows how a process or pattern is altered by a certain co-varying factor. It would be tragic to ignore or lose these new insights, as would happen if we adopted W2010's criteria.

In criterion 6, W2010 proposes a cutoff for the number of observations along an *x*-axis needed to differentiate linear from nonlinear relationships. We agree that one's ability to differentiate linear from nonlinear relationships can be an important issue when trying to understand productivity–diversity relationships. Detecting unimodality compared to a monotonically increasing relationship requires the occurrence of significant estimates of linear and quadratic regression terms, the detection of an internal maximum in the regression (e.g., Mitchell-Olds and Shaw 1987), and tests of whether the quadratic regression is a more parsimonious model than the simple linear one (e.g., by using Akaike information criteria, AIC, or log-likelihood methods; Burnham and Andersson 2002). However, we disagree with the need to impose some arbitrary cutoff for the number of observations needed to make a study useful. This criterion ignores the fact that meta-analytical techniques allow one to weight or reduce the impact of a study that is not well replicated, or to assign levels of confidence to studies that may be data poor (Gurevitch and Hedges 2001).

It becomes obvious from W2010 that the meta-analyses on productivity diversity relationships (PDR) by M2001 and others can be critically evaluated for including (or not including) certain studies or using certain methods. However, these problems remain unresolved by a very arbitrary list of "quality" criteria. Instead, the discussion should be by reanalysis of these data and the existence of this forum section reflects such a scientific progress. Although we share some points of criticism with W2010 on the lack of rigor in conducting these meta-analyses, we remain convinced that M2001 contributed much to the debate of PDR, as their analysis showed that the paradigm-like statement of a single universal hump-shaped PDR (Rosenzweig and Abramsky 1993) lacks empirical evidence. We do not expect a reanalysis of the M2001 database to change this general outcome.

## A General Critique for Meta-Analysis

Aside from the detailed arguments about the criteria that should be used to guide meta-analyses, our strongest point of dissent with W2010 is with his calling for a halt in conducting meta-analyses. It almost goes without saying that ecological data tend to be highly contingent on scaling issues, on seasonal and other intra-annual patterns, on inter-annual differences in abiotic constraints, on the type of experimental or observational approach, the chosen measure for a certain biological variable, and so on. But those who focus all their attention on such contingencies will inevitably miss the forest for the trees, and fail to see generality in ecological phenomena (Lawton 1999).

Meta-analyses are the remote-sensing tools of ecology. They allow us to step back from small-scale contingencies and see a broader, albeit less detailed, overview of how a system operates. A meta-analysis can give a baseline result for a certain process (e.g., the impact of grazing on plant biomass) to which new experimental studies can be compared. A meta-analysis can give a central tendency for a process, pattern or effect, which is debated in the literature and in cases show why results are different between studies. In the best cases, meta-analyses create new research hypotheses by showing what we do not know. It is immanent in this kind of analysis that peculiarities of certain ecosystems and organisms are not reflected. However, the goal of meta-analyses is to reveal pattern and process of the whole forest, not to show what's happening on the individual trees.

W2010 claims there are a number of technical shortcomings in three analyses of the PDR. His claims suggest that reevaluation and improvement of these meta-analyses might be useful. However, his suggestion that we halt meta-analyses is, in our opinion, short sighted. Not only does it neglect the power and usefulness of this tool, it ignores the many improvements of meta-analytical approaches achieved during the last decade. Ecologists have adopted different types of effect sizes (Osenberg et al. 1997, Gurevitch and Hedges 2001, LaJeunesse and Forbes 2003), have analyzed the statistical properties of these effect sizes (Hedges et al. 1999), and have improved their criteria for the inclusion of data (Englund et al. 1999). There is also increased awareness about the interdependency of data derived from one study and the importance of weighted meta-analyses. If these "best techniques" are not used correctly or reproducibly, then commenting on analyses and reanalyzing data is an integral part of the scientific process. However, calling for a halt in meta-analyses is like calling for cessation of cancer research simply because one drug didn't live up to everyone's expectations.

### Moving Beyond Pattern

W2010 discussed what he sees as limitations of the different meta-analyses on PDR. While we accept his argument that there have been shortcomings and flaws in meta-analyses that require a second look, we have disagreed with his vision for how synthesis via meta-analyses should proceed. We also believe that W2010 misses an important point in his comment that, in our opinion, is one of the primary limitations with research on PDR, that is, the lack of focus on the mechanisms that are presumed to generate PDR. With the possible exception of the debate over how diversity is related to stability (see Ives and Carpenter 2008), few discussions in ecology are in a worse state of understanding mechanisms than the discussion on PDR. There are numerous reasons for this. Here we discuss just two.

First, empiricists have used a plethora of different variables to represent "diversity" and "productivity." For example, consider that estimates of "productivity" range from variables as divergent as direct estimates of the rate of carbon flux through plants or animals, to the standing stock biomass of these same organisms, to the standing stock availability of resources used by these organisms, to the rates at which those resources are supplied, to highly derived covariates of resources or biomass such as latitude, depth, or elevation (M2001). Researchers often assume that the aforementioned variables are all mechanistically interchangeable, and that they operate consistently across varied trophic levels. Yet, the ecological theories on which predictions of PDR are often based suggest that species richness should be a function of (1) the supply rate of limiting resources that regulates species population sizes and stochastic rates of extinction (i.e., species–energy theory; see Wright 1983, Abrams 1995, Srivastava and Lawton 1998) and/or (2) the relative ratios of different resources that mediate competitive interactions and coexistence among species that share resources in a local community (resource-ratio theory; Tilman 1977, 1982). Empiricists tend to measure production and biomass as proxies for resource supply, which assumes there is a linear relationship between the availability of resources (what one might call the "potential" productivity of a system) and the conversion of those resources into new biomass (that is, the "actual" production of biomass). This may, at times and in some systems, be a legitimate assumption. However, it frequently is not: otherwise, why would we study things like Type-II functional response curves, compensatory feeding, assimilation efficiencies, and so on?

A second problem is that there is considerable confusion about whether productivity is the cause or the consequence of species diversity. Obviously, the supply rate of limiting resources, and the ratios at which different limiting resources are supplied, influence both the amount of biomass that can be achieved by a local community as well as the number of species it can support. As a result, species richness and productivity are often associated with one another. However, as discussed in the last paragraph, theory argues it is the supply rate of one or more resources, not productivity per se, that is the direct proximate cause of species richness. Plants don't generally consume or compete for their own tissue, and as such, theory doesn't predict a direct causal link from biomass or production of plants to species richness of plants. If anything, the causal link between richness and production goes in the opposite direction. Over the past two decades, there has been a wealth of experiments that have manipulated the richness of primary producers in terrestrial, marine, and freshwater ecosystems and shown that more species-rich communities capture limited inorganic resources more efficiently (reviewed in Balvanera et al. 2006, Cardinale et al. 2006). As a result, diverse communities tend to achieve higher biomass because species use limiting resources in ways that are complementary in space or time (see meta-analysis of Cardinale et al. 2007).

The contrast between the perspective that productivity-drives-diversity vs. the perspective that diversity-drives-productivity has led several authors to propose conceptual frameworks (Loreau et al. 2001, Schmid 2002, Cardinale et al. 2009) and mathematical models (Gross and Cardinale 2007) to explain how these views can be merged. These models share the common feature that the rates and/or ratios of resource supply (i.e., potential productivity) are what directly limit species richness in a local community. However, it is species richness that regulates the efficiency by which resources are captured and converted into new tissue. Importantly, these frameworks have also shown that when pathways of causality are mixed up, or biomass and resource supply are assumed to be interchangeable, one can observe spurious relationships between species richness and biomass that change as a function of spatial scale (Gross and Cardinale 2007). The key point here is that, if one is not careful to correctly identify the proximate causal and response variables, you can get a totally different picture of what the species richness–productivity relationship (SRPR) looks like.

Perhaps it is no surprise that ecologist have yet to produce a consensus view on the qualitative nature of SRPR. In our opinion, one of the primary contributions of M2001 was to illustrate the lack of a dominant and generalizable pattern of SRPR, which overturned a paradigm of a single unimodal PDR applying to all kinds of organisms and ecosystems (Rosenzweig and Abramsky 1993). This lack of generality almost certainly reflects to one degree or another the widespread use of proxies and lack of direct causal mechanisms linking the measured variables that have hampered our understanding.

If the M2001 analysis can be improved by quantifying studies more rigorously, then this would be a useful part of the normal scientific process But this does not justify suggesting to throw the baby out with the bathwater by halting meta-analyses on PDR or other important issues in ecology. Before any researcher undertakes a new synthesis of PDR, we hope s/he will think deeply about what direct proximate causal and response variables are involved in these relationships, and consider testing the improved conceptual frameworks that have been developed to help us better understand these relationships.

## Literature Cited

Abrams, P. A. 1995. Monotonic or unimodal diversity–productivity gradients: What does competition theory predict? Ecology 76:2019–2027.

Balvanera, P., A. B. Pfisterer, N. Buchmann, J. S. He, T. Nakashizuka, D. Raffaelli, and B. Schmid. 2006. Quantifying the evidence for biodiversity effects on ecosystem functioning and services. Ecology Letters 9:1146–1156.

Burnham, K. P., and D. R. Andersson. 2002. Model selection and multimodal inference. Second edition. Springer, New York, New York, USA.

Cardinale, B. J., H. Hillebrand, W. S. Harpole, K. Gross, and R. Ptacnik. 2009. Separating the influence of resource "availability" from resource "imbalance" on productivity–diversity relationships. Ecology Letters 12:475–487.

Cardinale, B. J., D. S. Srivastava, J. E. Duffy, J. P. Wright, A. L. Downing, M. Sankaran, and C. Jouseau. 2006. Effects of biodiversity on the functioning of trophic groups and ecosystems. Nature 443:989–992.

Cardinale, B. J., J. P. Wright, M. W. Cadotte, I. T. Carroll, A. Hector, D. S. Srivastava, M. Loreau, and J. J. Weis. 2007. Impacts of plant diversity on biomass production increase through time due to complementary resource use: a meta-analysis. Proceedings of the National Academy of Sciences USA 104:18123–18128.

Chase, J. M., and M. A. Leibold. 2002. Spatial scale dictates the productivity–biodiversity relationship. Nature 416:427–430.

Declerck, S., M. Vanderstukken, A. Pals, K. Muylaert, and L. de Meester. 2007. Plankton biodiversity along a gradient of productivity and its mediation by macrophytes. Ecology 88: 2199–2210.

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

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.

Foster, M. S., M. S. Edwards, D. C. Reed, D. R. Schiel, R. C. Zimmerman, M. A. Steele, S. C. Schroeter, R. C. Carpenter, D. J. Kushner, B. S. Halpern, K. Cottenie, and B. R. Broitman. 2006. Top-down vs. bottom-up effects in kelp forests. Science 313:1737–1739.

Gross, K., and B. J. Cardinale. 2007. Does species richness drive community production or vice versa? Reconciling historical and contemporary paradigms in competitive communities. American Naturalist 170:207–220.

Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. Ecology 80:1142–1149.

Gurevitch, J., and L. V. Hedges. 2001. Meta-analysis: combining the results of independent experiments. Pages 347–369 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Chapman and Hall, New York, New York, USA.

Gurevitch, J., and K. Mengersen. 2010. A statistical view of synthesizing patterns of species richness along productivity gradients: devils, forests, and trees. Ecology 91:2553–2560.

Gurevitch, J., L. L. Morrow, W. Alison, and J. S. Walsh. 1992. A meta-analysis of competition in field experiments. American Naturalist 140:539–572.

Halpern, B. S., K. Cottenie, and B. R. Broitman. 2006. Strong top-down control in southern California kelp forest ecosystems. Science 312:1230–1232.

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

Hillebrand, H., D. S. Gruner, E. T. Borer, M. E. S. Bracken, E. E. Cleland, J. J. Elser, W. S. Harpole, J. T. Ngai, E. W. Seabloom, J. B. Shurin, and J. E. Smith. 2007. Consumer versus resource control of producer diversity depends on ecosystem type and producer community structure. Proceedings of the National Academy of Sciences USA 104:10904–10909.

Holker, F., D. Beare, H. Dorner, A. di Natale, H.-J. Ratz, A. Temming, and J. Casey. 2007. Comment on "Impacts of Biodiversity Loss on Ocean Ecosystem Services." Science 316: 1285.

Ives, A. R., and S. R. Carpenter. 2008. Stability and diversity of ecosystems. Science 317:58–62.

Jaenike, J. 2007. Comment on "Impacts of Biodiversity Loss on Ocean Ecosystem Services." Science 316:1285.

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.

Lawton, J. H. 1999. Are there general laws in ecology? Oikos 84:177–192.

Loreau, M., S. Naeem, P. Inchausti, J. Bengtsson, J. P. Grime, A. Hector, D. U. Hooper, M. A. Huston, D. Raffaelli, B. Schmid, D. Tilman, and D. A. Wardle. 2001. Biodiversity and ecosystem functioning: current knowledge and future challenges. Science 294:804–808.

Mitchell-Olds, T., and R. G. Shaw. 1987. Regression analysis of natural selection: statistical influence and biological interpretation. Evolution 41:1149–1161.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381–2396.

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

Rosenberg, M. S. 2005. The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. Evolution 59:464–468.

Rosenzweig, M. L., and Z. Abramsky. 1993. How are diversity and productivity related? Pages 52–65 *in* R. E. Ricklefs and D. Schluter, editors. Species diversity in ecological communities. University of Chicago Press, Chicago, Illinois, USA.

Schmid, B. 2002. The species richness–productivity controversy. Trends in Ecology and Evolution 17:113–114.

Srivastava, D. S., and J. H. Lawton. 1998. Why more productive sites have more species: an experimental test of theory using tree-hole communities. American Naturalist 152:510–529.

Tilman, D. 1977. Resource competition between planktonic algae: an experimental and theoretical approach. Ecology 58: 338–348.

Tilman, D. 1982. Resource competition and community structure. Princeton University Press, Princeton, New Jersey, USA.

Whittaker, R. J. 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91:2522–2533.

Wilberg, M. J., and T. J. Miller. 2007. Comment on "Impacts of Biodiversity Loss on Ocean Ecosystem Services." Science 316:1285.

Worm, B., et al. 2006. Impacts of biodiversity loss on ocean ecosystem services. Science 314:787–790.

Wright, D. H. 1983. Species–energy theory: an extension of species–area theory. Oikos 41:496–506.

*Ecology*, 91(9), 2010, pp. 2550–2552

# Mega mistakes in meta-analyses: devil in the detail

Len N. Gillman[1,3] and Shane D. Wright[2]

[1]*School of Applied Science, AUT University, Private Bag 92006, Auckland, New Zealand*
[2]*School of Biological Sciences, University of Auckland, Private Bag, Auckland, New Zealand*

## Introduction

Science is a work in progress, with each new study attempting to add to, or improve upon, those that have come before. In this way we have moved initially from a characterization of the species richness-productivity relationship (SRPR) as being ubiquitously unimodel (Rosenzweig and Abramsky 1993, Huston and deAngelis 1994) to a transitional view of the relationship as one in which unimodel relationships were seen not to be dominant but to instead depend on the geographic scale of study (Mittelbach et al. 2001). More recently there has been a re-characterization in which the dominant form of the relationship has been found to be positive at both fine and course grains and at all but very local geographic scales (Gillman and Wright 2006). However, Robert Whittaker's (Whittaker 2010) analysis gives the impression that on this issue we have recently descended into chaos. He first suggests that a set of prescribed criteria should be followed, but then concludes that we should entirely abandon meta-analyses in favor of narrative review or more directed primary data collection. We have some sympathy with Whittaker's arguments but we take issue with the analysis he has undertaken and the conclusions he makes. We revisit his analysis and in so doing conclude that he has overstated the problem and that the way forward is not to abandon meta-analyses, but to ensure that greater caution is exercised when undertaking, reviewing and citing them. The habitual problem with any meta-analysis is, we believe, not primarily with the statistical analysis, but with the widespread indiscriminate use of studies that are fed into the analysis, that are entirely inappropriate to the question being asked. Statistical meta-analysis can resolve some of these issues. However, in some cases the statistical analysis that has been performed, rather than overcoming problems, adds to the deception with a veneer of respectability. Unfortunately, poorly derived meta-analyses continue to be cited without question.

## Selection of Studies for Meta-analyses

Whittaker (2010) suggests that narrative review is a preferable option to that of meta-analysis for summarizing the empirical literature. However, this suggestion ignores the fact that such reviews involve author selection and interpretation, almost always without the use of formal selection criteria, and are therefore not immune from the same problems that can produce misleading meta-analyses. Meta-analysis was born out of a need for more objective assessments of large, apparently conflicting, bodies of empirical study than can be reasonably managed by narrative review. Moreover, as the number of studies increases, so does the influence of type II error and as a consequence sound hypotheses become, ironically, more vulnerable to false rejection as the body of empirical science increases (Hunter and Schmidt 2004).

Where we are in agreement with Whittaker (2010) is in his concern for the general lack of scrutiny applied to the selection of data sets for meta-analysis. However, on this issue, there is some controversy (Gurevitch and Hedges 2001). There are two potentially valid approaches to this problem: (1) to apply selection criteria based on "good a priori evidence" about factors that bias study results (Englund et al. 1999, Hunter and Schmidt 2004); (2) to include all studies and to tests hypotheses within the meta-analysis relating to possible factors that might affect the results. If the apparently "poor-quality" studies do not produce different results than the "good-quality" studies, then there is no justification for excluding them (Englund et al. 1999, Gurevitch and Hedges 2001, Hunter and Schmidt 2004). The latter approach is favored by many because the former potentially suffers more from author bias, as demonstrated by Englund et al. (1999), and can lead to unnecessary reduction in the sample size. If the first option is used "it is critical that the description of data selection is explicit" (Englund et al. 1999:1134) and "the criteria for inclusion should be reasonable and scientifically defensible" (Gurevitch and Hedges 2001:352).

The problem, highlighted by Whittaker (2010), arises when neither of these approaches are followed such that studies are, on the one hand, included that suffer from fatal experimental design faults and are then counted

Manuscript received 26 February 2009; revised 12 May 2009; accepted 15 May 2009. Corresponding Editor: T. J. Stohlgren. For reprints of this Forum, see footnote 1, p. 2534.

[3] E-mail: len.gillman@aut.ac.nz

with equal weight to those that use more appropriate study designs. On the other hand, other studies may be excluded without reference to specific and defensible criteria. Many of the studies, for example, included by Mittelbach et al. (2001) and Partel et al. (2007) used rainfall as a surrogate for productivity, but confounded increasing rainfall with increasing elevation where depressed temperatures limit productivity. Further, they used soil nutrients as a surrogate for productivity but confounded nutrient status with increasing salinity which is toxic to most species. Other studies accepted by the above authors used quadrat sizes that differed within the study or that were too small to accommodate whole plants. The problem with these studies arises from factors that are well known to influence either the predictor variable (productivity) or the response variable (species richness) and are likely to have seriously influenced the apparent form of the productivity-species richness relationship. Therefore, the inclusion of such studies, without further analysis of their influence, does not present a valid characterization of the relationship. Unfortunately, the invalid conclusions based on such studies continue to be cited, as demonstrated in Table 1 of Whittaker (2010).

Whittaker (2010) suggests seven criteria that he deems necessary for including data sets in plant species richness–productivity meta-analyses. We previously employed all but the first of these (Gillman and Wright 2006:1235–1237) along with the additional criterion that the sampling grain could not be less than the space likely to be occupied by single plants in the vegetation types included in the member study. Most of these criteria are based on good a priori evidence that particular factors will invalidate the inclusion of some studies. However, for example, those involving minimum sample size or the type of measure of diversity (e.g., species richness vs. genus richness), may or may not influence the results. The preferable approach, therefore for assessing quality issues would be to include all studies in the initial analysis and to follow this with an examination of how various quality factors affect the outcome. In this way, a more informed and transparent analysis can be presented.

## Devil in the Detail

Whittaker (2010) points out that despite critical reanalysis of the Mittelbach et al. (2001) study by Whittaker and Heegaard (2003) and ourselves (Gillman and Wright 2006), it is the former study that is preferentially cited in the literature and that recent meta-analyses have reused the data-base of Mittelbach et al. (2001) without paying heed to the problems with that data. We agree with Whittaker (2010) in as much as the advancement in knowledge is failing to gain traction in this instance and, in the vacuum, an apparently respectable but flawed characterization of species richness patterns endures.

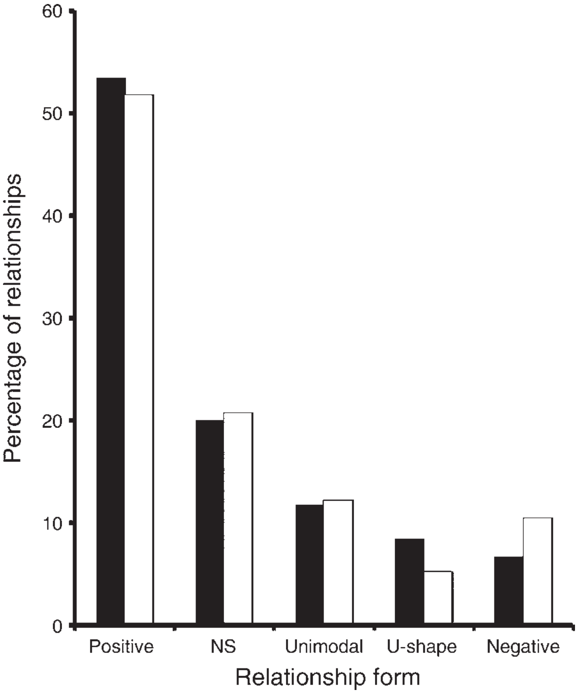

Fig. 1. Productivity–species richness relationships as published by Gillman and Wright (2006; solid bars, $N = 60$) and as adjusted if all differences with Whittaker (2010) were to be accepted (open bars, $N = 58$).

Whittaker (2010) bases his argument to abandon all meta-analyses on a review of 68 data sets used by one or more of the four studies (Mittelbach et al. 2001, Whittaker and Heegaard 2003, Gillman and Wright 2006, Partel et al. 2007). He claims his review of the case studies demonstrates considerable inconsistencies among the studies, with only 11% agreement among all four studies. However, closer examination of these case studies reveals that, although Whittaker (2010) disagrees with the assessments of Mittelbach (2001) in 24 cases and with those of Partel et al. (2007) in 30 cases, there is no inconsistency between Whittaker and Heegaard (2003) and our study (Gillman and Wright 2006) and only in four cases does Whittaker (2010) draw a different conclusion from ours. Two of these differences are due to Whittaker reclassifying studies on the basis of a visual purview of the data. A third difference is regrettably due to an error in our data set where the same source data was double counted from two publications. However, the other mistake that Whittaker points to in our study was not our mistake, but Whittaker's. Whittaker states that we classified the Wheeler and Shaw (1991) study "as U-shaped, and claim incorrectly that M2001 did the same." However, *Ecological Archives* (E082-024) clearly shows that Mittelbach et al. (2001) did classify this study as U-shaped.

The important point is, however, that the differences between our results and those of Whittaker's are somewhat trivial. If we were to accept all four reevaluations by Whittaker (2010) that differed from ours, it would make no material difference to the results we reported in 2006. Positive relationships remain dominant (Fig. 1). The discrepancies identified by Whittaker (2010), therefore, largely occur between Mittelbach et al. (2001) and Partel et al. (2007) on the one hand, and Whittaker and Heegaard (2003) and us on the other. We therefore suggest that the inconsistencies highlighted by Whittaker (2010) are largely due to Partel et al. (2007) following Mittelbach et al. (2001) without paying head to the limitations that have become apparent in the dataset of the latter study. We do not believe that a call to abandon all meta-analyses largely on the basis of one such study is justified.

## Conclusion

The perilous state of meta-analyses portrayed by Whittaker (2010) is, in our view, an overstatement of what is nonetheless a regrettable misuse of published data sets. Nearly all of the discrepancies among studies highlighted by Whittaker (2010) are due to the indiscriminate use of studies by Mittelbach et al. (2001) and Partel et al. (2007). One or two such studies, however, do not warrant the call for abandonment of all meta-analyses. Nor is the use of prescribed criteria necessarily the best method to overcome these problems. The alternative of including all studies and testing hypotheses relating to the confounding effects of including particular types of study is regarded by many as a preferable approach. However, the current situation where studies are included within meta-analyses without scrutiny or analysis for biasing influences is clearly creating a disservice to the discipline of ecology. We suggest the way forward is not to abandon meta-analysis but instead for researchers to apply greater care in constructing and analyzing them. There is also an imperative of greater care required in interpreting and citing meta-analyses that neither, employ defensible selection criteria, or undertake adequate post hoc analysis of potentially confounding effects of member studies. We hope that the spotlight placed on this issue by this forum will result in better practices being adopted in the future.

## Literature Cited

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.

Gillman, L. N., and S. D. Wright. 2006. The influence of productivity on the species richness of plants: a critical assessment. Ecology 87:1234–1243.

Gurevitch, J., and L. V. Hedges. 2001. Meta-analysis: combining the results of independent experiments. Pages 347–369 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Oxford University Press, Oxford, UK.

Hunter, J. E., and F. L. Schmidt. 2004. Methods of meta-analysis: correcting error and bias in research findings. Sage Publications, London, UK.

Huston, M. A., and D. L. deAngelis. 1994. Competition and coexistence: the effects of resource transport and supply rates. American Naturalist 144:954–977.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381–2396.

Partel, M., L. Laanisto, and M. Zobel. 2007. Contrasting plant productivity–diversity relationships across latitude: the role of evolutionary history. Ecology 88:1091–1097.

Rosenzweig, M. L., and Z. Abramsky. 1993. How are diversity and productivity related? Pages 52–65 *in* R. E. Ricklefs and D. Schluter, editors. Species diversity in ecological communities. University of Chicago Press, Chicago, Illinois, USA.

Wheeler, B. D., and S. C. Shaw. 1991. Above-ground crop mass and species richness of the principal types of herbaceous rich-fen vegetation of lowland England and Wales. Journal of Ecology 79:285–301.

Whittaker, R. J. 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91:2522–2533.

Whittaker, R. J., and E. Heegaard. 2003. What is the observed relationship between species richness and productivity? Comment. Ecology 84:3384–3390.

*Ecology*, 91(9), 2010, pp. 2553–2560
© 2010 by the Ecological Society of America

# A statistical view of synthesizing patterns of species richness along productivity gradients: devils, forests, and trees

Jessica Gurevitch[1,3] and Kerrie Mengersen[2]

[1]*Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York 11794-5245 USA*
[2]*School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia*

## Introduction

Robert Whittaker (2010) offers a critique of quantitative research syntheses attempting to generalize species richness patterns along gradients of productivity. As he says, the results of such syntheses have been controversial and disagreed in their conclusions. Beginning with a large research synthesis by Mittelbach et al. (2001), a number of authors (Gillman and Wright 2006, Pärtel et al. 2007, Laanisto et al. 2008) have attempted to classify patterns from individual studies by the shape of the responses; Whittaker and Heegaard (2003) criticized what they felt were methodological flaws in the Mittelbach et al. (2001) paper and the critique was rebutted by Mittelbach et al. (2003). Due to what he feels are persistent methodological flaws in the papers attempting quantitative syntheses of this literature, and because he now believes that it is impossible to carry out meaningful meta-analyses on this relationship, Whittaker (2010) recommends an end to meta-analyses on this topic, expresses concern over whether quantitative data synthesis is a legitimate and repeatable approach to making sense of the data on this question, and suggests a profound change in the way meta-analyses are conducted and reviewed in ecology.

We offer a short discussion of why we feel that Whittaker's dismissal of meta-analysis is inappropriate, add a brief critique of this literature of our own from a statistical perspective, and most importantly, point the way to improved statistical approaches. While the devil is certainly in the details, we also don't want to lose sight of the forest for the trees.

While Whittaker is correct in demanding unambiguous and transparent criteria for selecting studies and carrying out the meta-analysis, this is not a legitimate argument against the use of meta-analysis. In fact, because there are established criteria for carrying out systematic reviews and quantitative research syntheses—a broader field in which meta-analysis is one component—meta-analyses that follow contemporary established protocols are more likely to be repeatable than other forms of literature reviews (e.g., Borenstein et al. 2009). Whittaker is among many authors who have suggested specific criteria for study inclusion in ecological research synthesis; for example, Hillebrand and Cardinale (2010) argue for using different criteria. Just as the past four decades of debate about the process of systematic review in medicine have seen the establishment of widely accepted standards and protocols by the Cochrane Collaboration, so this engaged discussion can contribute positively to the development of the application of scientific principles and methodology to research synthesis in ecology (description of the Cochrane Collaborations *available online*).[4] More broadly, various formal techniques have been applied to account for variation in study quality in meta-analysis, including weighting by both inverse variance and study quality, or including moderators to account for methodological flaws in the statistical models employed (e.g., Cooper et al. 2009). Methods for dealing with variation in selection criteria among meta-analyses have been developed focusing on levels of generalizability of the outcomes (Sutton et al. 2000, Wolpert and Mengersen 2004; B. J. Becker and A. Aloe, *unpublished data*). Similar problems are endemic to research review and synthesis regardless of the method of synthesis that is adopted. For example, they are just as much a problem in qualitative evaluation as they are in quantitative pooling of effect sizes. A model-based approach to meta-analysis allows for the possibility of formal, transparent adjustment for selection bias and similar problems, whereas this is much more difficult with narrative methods or vote counts.

A second reason that we do not find Whittaker's dismissal of meta-analysis convincing is that flaws in particular research syntheses, real or perceived, do not invalidate meta-analysis itself. By carrying out a systematic review or quantitative research synthesis using flawed aims, assumptions or data, one may indeed undermine the validity of its conclusions. However, this is true for any scientific enterprise or statistical technique (Gillman and Wright 2010, Hillebrand and Cardinale

Manuscript received 10 June 2009; revised 1 September 2009; accepted 2 September 2009. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 2534.

[3] E-mail: Jessica.gurevitch@stonybrook.edu

[4] ⟨www.cochrane.org⟩

2010). We would not conclude that we should dismiss all future attempts at ecological lab or field work because of individual flawed studies, and likewise, while regression and other statistical techniques in ecology are often applied incorrectly and inappropriately, we would be reluctant to discard their use on that basis. Otherwise, we would be back to narrative case studies and individual natural history observations, as were standard fare until the 1960s when modern statistical techniques first became widely adopted in this field (Simpson et al. 1960).

Similar discussions occurred in medicine over a decade ago (see review and discussion, e.g., by Borenstein et al. 2009:384–386). Different issues regarding the adoption of meta-analysis methodology have been raised in different disciplines. For instance, in medicine, there were fierce debates about the use of fixed vs. random effects models in meta-analysis; this has not been an issue at all in the ecological literature. On the other hand, a number of key similarities exist in the arguments for and against the incorporation of scientific principles for research review (including meta-analysis) across disciplines. Issues such as how to define and account for variation in study quality in meta-analysis methodology, for instance, are common in all disciplines. Lau et al. (*in press*) explore the comparisons between the history of the adoption of meta-analysis in medicine, social scientific research, and ecology, and the parallels and contrasts are enlightening.

Whittaker (2010) suggests that more appropriate solutions than meta-analysis to understanding the diversity-productivity relationship are to devise rigorous field studies that will make meaningful contributions to the question, or to undertake a narrative review. Devising more field studies may be useful for filling in knowledge gaps, improving the quality of the available data and for many other reasons, but it will not help synthesize the results of the existing studies. Moreover, no one study can replace all of the existing studies or be complete enough to resolve this question across all systems, organisms, and scales. Interestingly, this parallels a debate in medicine, where the relative value of research syntheses and very large clinical trials with tens of thousands of subjects and sometimes lasting many years has been discussed at length; e.g., Lau et al. (1992), LeLorier et al. (1997), and Ioannidis et al. (1998). Nor will narrative reviews provide closure on this question. If a narrative review seeks generalizations, it will face most of the same problems that Whittaker (2010) finds with quantitative reviews (and others, besides, including reviewer bias and vote counting; e.g., Lipsey and Wilson 1993, Sharpe 1997). On the other hand, we stand to gain little from a narrative review that discusses each study as a unique example that is not comparable to others and where there is no generalization possible. In fact, if it is utterly impossible to generalize from the results of a study, or to compare studies to reach generalizations, we would argue that there is little value in the individual studies as well, because the exact circumstances of any one ecological study are unlikely to ever occur again.

Whittaker's final recommendation is to use "best evidence synthesis" (Slavin 1995). Slavin's ideas have been developed considerably since the publication of that and an earlier paper (Slavin 1986). Slavin's recommendations for clear problem statement, formalized literature search, critical literature review, evidence tabulation, and qualitative synthesis are now part of (albeit with some controversy) the larger body of work on systematic reviewing, which we cannot discuss at any length here, and many of these elements are standard practice in high quality meta-analyses. With the exception of qualitative synthesis, which is prone to biases and may indeed be less transparent than the quantitative approaches, these activities should be part of any synthesis, whether it is nominally "best evidence" or "meta-analysis." The adoption of these methods and their incorporation into current practice is an example of substantive progress in the establishment of systematic criteria for literature review.

As pointed out by Hillebrand and Cardinale (2010) and by many others, meta-analysis does not necessarily require that the aims of the original studies be the same, that there are no modifying factors that differ among studies, or that sampling schemes and study designs be identical among studies, as claimed by Whittaker (2010). These are neither conceptual nor statistical requirements for quantitative data synthesis, and rigorous statistical methodology has been developed to deal with all of these issues. Modifying factors that influence the outcomes and that vary among studies can in many cases be modeled. Of course, if there is true confounding among these factors, this can limit the inferences possible. While none of these issues are uniquely problematic for meta-analysis, unlike meta-analysis, other methods (e.g., narrative reviews) have not developed methods for addressing them. It is, of course, essential that studies be synthesized in a meaningful manner, and this can be challenging.

## Limitations to Vote Count Approaches

Each of the quantitative research syntheses on the productivity–diversity relationship from Mittelbach et al. (2001) on proceed by determining the shape of the curve in each of the primary studies being synthesized according to various criteria, and then tallying the numbers in each shape category and comparing them. Due to different criteria (and other factors) they arrive at different numbers of unimodal, U-shaped, and positive and negative monotonic curves across the literature. A fundamental problem with this approach is that the results are obtained using a statistical technique known as vote counting (Hedges and Olkin 1980, Gurevitch and Hedges 1999, Borenstein et al. 2009), in which studies are judged by their significance levels to "cast a vote" in favor of or against a particular outcome. Vote counting involves simple estimation of the proportion of studies that show a "significant" effect (where the definition of statistical significance may vary from one reviewer to another) in response to a specified hypothesis. In the species richness and productivity assessment, the primary test is whether the quadratic coefficient is not equal to zero (i.e., there is a curvilinear relationship).

The disagreement and confusion found among these studies is precisely what one would predict from a set of vote counts, because it is a flawed statistical technique that results in biased and inconsistent outcomes when used as a research synthesis method. Unfortunately vote counting is not only a weak form of inference, it is potentially misleading and may not provide the answers to the questions authors are generally most interested in addressing when synthesizing results across studies: the overall magnitude and direction of a parameter or an effect (such as a slope) and explanations for variation in that effect. In the case of the diversity–productivity relationship, there are various potentially important covariates and confounders, including the issue of scale. Because of the statistical problems with vote counts, the practice of applying simple or elaborate statistical tests to vote counts are likely to be misleading as well. Vote counting has essentially been abandoned in other disciplines but continues to be relied upon in ecology. Vote counting is not meta-analysis although it is sometimes misidentified as such in the ecological literature. At least some of the arguments made by Whittaker (2010) are valid critiques of vote counting, rather than meta-analysis.

Some of the reasons given by ecologists for adopting vote counting are that the data are not available for doing a meta-analysis, that the results are too heterogeneous to warrant a formal meta-analysis, and that vote counts are somehow more conservative or reliable than meta-analysis because they appear to have fewer assumptions (e.g., Ives and Carpenter 2007, Tylianakis et al. 2008).

In fact, like any statistical technique, vote counts also are based on assumptions, although these are often not examined. Missing data (e.g., means, sample sizes, and errors) can indeed pose major challenges for meta-analysis. Methods have been proposed to deal with partial missing data (e.g., Fahrbach 2001, Pigott 2009; Lajeunesse and Schmid, *in press*) but if too much data are unreported it may be impossible to accurately synthesize the data quantitatively; vote counts will not provide more precise or accurate syntheses in this case. If an effect is consistently reported over different scales, locations, time periods, and study designs, a vote count may provide some support for a true association. However, the analyst should clearly state the reasons for the adoption of vote counting despite its limitations and the limited inferences that can be made on the basis of such an analysis. Vote counts may also prove to be of some use in a first-pass exploratory data analysis, to gauge the overall patterns of responses. Another alternative where the analyst feels that data are too limited for formal meta-analysis is combining $t$ statistics. Like vote counting and combining $P$ values, the combination of $t$ statistics can accommodate heterogeneous study results, but unlike vote counting or combining $P$ values, it has the advantage of taking into account the magnitude of the study-specific effect estimates. This approach is an improvement on the combination of $P$ values, since the latter does not discriminate between positive and negative values, but still suffers from other major drawbacks (e.g., see Becker and Wu 2007).

In the case of the productivity–diversity relationship, the results are not consistent across studies, but vote counting is not a good tool for analyzing the sources of this heterogeneity. If the results of a group of studies are strongly heterogeneous or cannot be statistically combined for other reasons, one option might be to conduct a descriptive narrative review that does not depend on the $P$ values of the outcomes of the studies being combined. A review that categorizes studies by characteristics other than those based on the statistical significance of the outcomes may be informative and meaningful; for instance, a review that finds that 70% of studies on invasive species concern only plants is an interesting and useful finding and is not subject to the limitations of vote counts based on significance tests. If a researcher deems that a group of studies is irreducibly heterogeneous at all levels of generality, it may be worthwhile to reframe the research question to something more meaningful and tractable.

The charge of misleading inferences arising from vote counting follows in part from two major drawbacks of this approach: it takes no account of the magnitude of the effect or of the uncertainty in the estimate of that effect (i.e., the confidence interval around the effect estimate). As a very simple example, if vote counting of "statistically significant" studies is used to determine if an effect is substantiated across studies, and if one study shows a very strong quadratic relationship and two studies show a "nonsignificant" quadratic effect, vote counting would, possibly erroneously, lead to a conclusion of "inconsistent effects" or of "no overall effect." Similar problems exist if counts of reported (or computed) "significant" U-shaped or hump-shaped relationships are compared. Moreover, a change in the threshold for tests of "significance" can subtly or dramatically change the "votes" and thus the resultant inferences. Mittelbach et al. (2003:3393) acknowledged this issue, stating that "the strength of the quadratic terms is a legitimate issue separate from its existence and this is not something we attempted to address... ." More formally, the vote count estimate does not meet any of the criteria for a good estimator—it is not unbiased, consistent, or sufficient—so its usefulness in providing meaningful quantitative syntheses is quite limited.

## Statistical Approaches to Quantitative Synthesis of the Diversity–Productivity Relationship

What are the alternatives to vote counting? Contemporary meta-analysis practice relies upon on a model-based combination of the study-specific data. In brief, meta-analysis involves obtaining a measure of the effect from each study, called an "effect size" (such as a standardized mean difference, response ratio, correlation coefficient, or a regression coefficient), weighting the effect sizes by the inverse of their sampling variances, and then modeling these weighted mean outcomes across studies. We note that there are substantial benefits to weighting in this way, which is why it has become standard practice in meta-analysis. These include being able to model the within- and between-study heterogeneity, and accounting for differences among studies in the precision of the effect size estimate. Weighting effect estimates by their respective (inverse) variances has mathematical properties that may be lost if the weights are formed using some other, arbitrary criteria.

Ecologists have occasionally objected to using variances in meta-analysis either as a basis for effect size calculations (e.g., as used in standardized mean differences) or as weights, with the rationale that there may be systematic differences between field and lab studies in the magnitude of variances, creating a statistical bias in favor of lab-based studies. Curiously this hypothesis (that lab-based studies have smaller variances, or larger effect sizes on average than field-based studies), while not unreasonable, has never been demonstrated to be true. Moreover, the overall effect estimates that are obtained with arbitrary weights must be auditable (i.e., based on a robust mathematical justification for selection of the weights) and must be interpretable. This is still a matter of great debate in the statistical literature. In any case, if there are systematic differences such that bias is introduced by combining two very different types of data, the synthesist certainly has the option of analyzing those two groups of studies separately rather than combining them. This is a better alternative than discarding a valuable statistical tool for which there is no obvious substitute.

The techniques for modeling the outcomes (i.e., effect sizes) across studies have experienced considerable development over the past four decades, and can range from the very simple—e.g., finding a weighted grand mean across studies and its confidence limits—to modeling variation in the effects across studies, including both frequentist and Bayesian approaches (e.g., Hedges and Olkin 1985, Borenstein et al. 2009, Cooper et al. 2009). These techniques offer statistically unbiased and robust means for asking questions about the overall magnitude and direction of the effect and about heterogeneity among studies, i.e., variation in the magnitude, confidence limits (or credible intervals), and statistical significance of that effect.

Let us assume that the meta-analyst has compiled a set of studies that are biologically relevant, satisfy conditions of comparability of scale, meet study quality and design thresholds, report consistently on important covariates (that is, factors influencing the nature of the relationship between productivity and diversity), and provide either primary or summary data about the relationship of interest. In the present case, this relationship is the association between species richness and productivity; the primary data available from each study would comprise a set of pairwise values of the two variables over a defined range This is one of the approaches taken by Mittelbach et al. (2001), who obtained relevant data from the 171 studies published studies and then reanalyzed the available study-specific data using well-defined linear and quadratic regression models in a formal, if limited, meta-analysis. Alternatively, the summary data from each paper would be the regression parameter estimates (intercept, linear coefficient, quadratic coefficient, and so on), the corresponding standard errors of these estimates and/or the $t$ or $P$ values resulting from the hypothesis that the parameter estimates are equal to zero. An interesting recent hybrid between a very large primary study and the use of study-specific data synthesis on the productivity–diversity relationship in marine systems was recently published by Witman et al. (2008).

The compilation of primary data from all studies is often argued to be a "gold standard" approach in meta-analysis. One important benefit is that synthesis based on primary data allows for consistent analysis of data within studies and thus provides a directly comparable set of study-specific estimates for input into a meta-analysis. There are also drawbacks to obtaining and analyzing study-specific data: the process is extremely time-consuming and it is often not possible to obtain the required data from all identified studies. It may be possible to extract data from published figures if the original data are not available, although this may potentially introduce inaccuracies (Whittaker and Heegaard 2003). Although the ecological community is beginning to redress the problems of poor data reporting with increasing pressure on authors to make complete data sets available online, they will persist for the foreseeable future since meta-analysis relies on a historical profile of published papers. A global problem is the possibility of incorrectly analyzing data due to ignorance of important characteristics of the study design and conduct, adjustment for biases and confounders, data collection and management, treatment of missing outcomes and covariates, and so on.

A strong advantage of having access to the primary data is the ability to build more comprehensive meta-analysis models. In the meta-analysis of quadratic (or other order) regressions, a number of options are possible. For example, the meta-analyst can build a hierarchical (multilevel) model that allows for separate study-specific regression models at the local level, and then combines the regression parameters using fixed- or random-effects assumptions at a global level. Intermediate levels can be included to describe subgroups of studies with common parameter values. Texts describing these methods include Raudenbush and Bryk (2002), Goldstein (2003), and West et al. (2007), and papers on their application in ecology include Helser and Lai (2004) and Thompson and Hobbs (2006). Alternatively, a meta-analysis analog to an ANOVA approach can be taken, whereby individual regressions are fitted to each study, then a test is conducted to assess whether a common quadratic coefficient can be fitted, which if passed is followed by a test to assess whether a common linear coefficient can be fitted, and finally whether a common intercept can be fitted (Tweedie and Mengersen 1995). In both of the above cases, the analysis can include an assessment of the contribution of the higher order polynomial terms; that is, the increased goodness of fit achieved by including a quadratic term at all is formally tested, which if failed is followed by a test of the linear term.

If the meta-analyst uses the primary data to obtain summary estimates of the regression parameters, the resultant data set is equivalent to that obtained by extracting these summary estimates from the published papers themselves (assuming that the same models have been fitted and that the required information is published or otherwise available). The meta-analyst now has a choice between a simple combination of $t$ statistics or a more complete approach involving combining parameter estimates according to an explicit statistical model. In the study of the association between species richness and productivity, these parameter estimates might include correlations, linear and quadratic coefficients, and goodness of fit measures.

This brings us, then, to the more complete statistical approach to combining the parameter estimates themselves. In the context of a quadratic regression, this involves fitting a (typically) random effects model to the multiple estimates of intercept, linear coefficient and quadratic coefficient. A random effects model is generally preferable because we would usually assume in ecology that the regression coefficients differ among studies, in addition to which there is sampling error between the estimates of the coefficients among studies.

Thus, assume that the fitted model for the $j$th study is

$$\hat{\mathbf{y}}_j = b_{0j} + b_{\mathrm{L}j}\mathbf{X}_j + b_{\mathrm{Q}j}\mathbf{X}_j^2$$

where $\hat{\mathbf{y}}_j$ is the vector of expected responses and $\mathbf{X}_j$ is the vector of covariates. This model can be fitted to the original or transformed values for the study outcomes (e.g., log transformation) and under different assumptions may be represented and estimated using ordinary least squares or a variant of this approach (iterative least squares, generalized least squares, etc.), or as a generalized linear model. Assuming that the regression coefficients are normally distributed, a simple random effects model that then combines these study-specific regression estimates is as follows:

$$\begin{pmatrix} b_0 \\ b_{\mathrm{L}} \\ b_{\mathrm{Q}} \end{pmatrix}_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{C}_j)$$

$$\boldsymbol{\mu} = \begin{pmatrix} \beta_0 \\ \beta_{\mathrm{L}} \\ \beta_{\mathrm{Q}} \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_0^2 & \sigma_{0\mathrm{L}} & \sigma_{0\mathrm{Q}} \\ \sigma_{0\mathrm{L}} & \sigma_{\mathrm{L}}^2 & \sigma_{\mathrm{LQ}} \\ \sigma_{0\mathrm{Q}} & \sigma_{\mathrm{LQ}} & \sigma_{\mathrm{Q}}^2 \end{pmatrix}$$

where $\boldsymbol{\mu}$ is the global mean vector of regression estimates, $\boldsymbol{\Sigma}$ is the between-study variance–covariance matrix, and $\mathbf{C}_j$ is the study-specific variance–covariance matrix for the parameter estimates (the intercept, linear, and quadratic coefficients). In the context of estimating the shapes of the productivity–diversity curves across studies, the estimate of the overall quadratic regression estimate, $\beta_{\mathrm{Q}}$, is perhaps of primary interest. This model is described by Becker and Wu (2007) in a general statistical context and by Arends (2006) as one of a wide range of methods for multivariate meta-analysis. On the other hand, if one is only interested in the existence (and magnitude) of a quadratic effect and not the whole shape of the productivity–species richness relationship, then it might suffice to undertake a univariate (random effects) meta-analysis on the quadratic coefficients, rather than the full multivariate model (comprising the combination of intercept and linear and quadratic coefficients).

Note that the above model explicitly accommodates both within-study variation and between-study heterogeneity, which is possible because the individual study parameters are weighted by the inverse sampling variances. Specific sources of heterogeneity can be included as covariates by directly extending the above model to a meta-regression, or through a multi-level model in which different subsets of the studies are combined within and across the different levels. Heterogeneity is dealt with as in any other meta-analysis: if it can be described using covariates (moderators), one can extend the model to a meta-regression; additional variation can be included either as simple between-study variation using a random-effects model, or more elaborately in a hierarchical model. This proposed statistical approach may not be applicable if the study designs are very different, if insufficient data are available to calculate the parameters, or if the responses are on very different scales, among other limitations.

Mittelbach et al. (2001) and others discuss the problem of determining whether a putative maximum or minimum is indeed a turning point within the observed range of productivities, that is, whether the regression warrants a quadratic rather than linear form. In a quantitative meta-analysis such as the model described above, all of the quadratic coefficients, whether "U shaped" or "hump shaped" or "no maximum or minimum within the range studied," can be combined to provide an overall estimate of the species richness–productivity relationship. This is the same as combining positive and negative estimates of any measure in a meta-analysis and then testing for heterogeneity among the studies, and depends of course on the ecological validity and interpretation of the result. If desired, tests such as the Mitchell-Olds and Shaw (1987) test applied by Mittelbach et al. (2001) can then be applied to assess whether this overall estimate is a turning point within the observed range of productivities. If the results are heterogeneous, a hierarchical statistical model could be explored to test for differences in the shape among groups of studies, and for heterogeneity within groups. Alternatively, in a Bayesian framework, we suggest that the assessment (max/min/neither) might be embedded in the analysis, by counting the number of times in an MCMC simulation that the turning point was in the appropriate productivity range and, if so, whether the quadratic term was negative (giving the probability that the relationship is "hump shaped") or positive (giving the probability that the relationship is "U shaped"). The same analysis can provide separate estimates of the positive and negative relationships if desired. Based on the above model, these probabilities can be estimated for each study as well as for the overall relationship.

## Discussion and Conclusions

The general model for meta-analysis of regression coefficients described above is well established in the literature. Jones et al. (1994) describe an early example of its use for the meta-analysis of 42 published experiments of mitochondrial electron transport; here, nonlinear regression was used to estimate the relationship of interest in each study and the results of the regression analyses were synthesized by a random effects model. Van Houwelingen et al. (2002) provide a general description of the bivariate version of this model (i.e., intercept and linear term only), and Paul et al. (2006) applied this in an ecological context, with the aim of analyzing the relationship between fusarium head blight and deoxynivalenol content of wheat among 126 field studies. A Bayesian analogue of the model has also been described by Riley et al. (2007). The same model framework can be used to combine other parameters of interest such as correlations; see Paul et al. (2006) for an example and discussion.

More flexible regression models, such as fractional polynomial regression and spline regression, may also be considered as alternatives to quadratic (and higher-order polynomial) descriptions of a nonlinear relationship. These models can be applied to the study-specific data and then combined using an inverse-variance-weighted random-effects model; Bagnardi et al. (2004) describe this approach in an epidemiological context. To our knowledge, this has not been applied to any ecological problems; it may also be interesting to assess whether this would assist with the problem of asserting that a maximum/minimum has occurred in a specified range.

We note that Mittelbach et al. (2001) did utilize the multivariate meta-analysis model described above and obtained a negative parameter estimate for the overall quadratic coefficient, $\beta_Q$, with an associated 95% confidence interval that did not include zero. However, the authors embedded this result in the body of the paper, preferring to use it as a supplementary rather than primary analysis, and did not elaborate on it more than obtaining the overall mean, confidence limits and heterogeneity. Not surprisingly, the quadratic coefficient was highly heterogeneous across studies, but this was not explored further quantitatively.

A drawback of the model described above is the need for estimates of the covariances of the study-specific regression parameters. These are rarely reported in the published papers and are typically difficult or impossible to estimate using surrogate information. Thus adoption of this method usually relies on access to the primary data. However, alternatives are being devised. For example, Riley et al. (2008) suggest a slight reparameterization of the random effects meta-analysis model that does not involve the covariance terms, at the expense of modified inferences. Although the synthesis of regression parameters was not explicitly discussed, it may be possible to transfer these results to this context.

We recognize that the model expressed in the form above does not explicitly address many of the substantive issues identified by Mittelbach et al. (2001), Whittaker and Heegaard (2003), Gillman and Wright (2010), and others. For example, if the gradients in the individual studies are not of the same length, it is probably meaningless to combine parameters across all studies; instead, the above model could be applied to ecologically more homogeneous categories of studies, such as those identified by Mittelbach et al. (2001) based on local and global scale ranges. Moreover, instead of maintaining these as separate analyses, an additional hierarchy could be added to the model that allows combination of the outputs from the different categories; again, although this is valid statistically, the resultant estimates and inferences would need to be ecologically interpretable and supportable. If plot size was considered to be the most ecologically meaningful covariate (Whittaker 2010), similar approaches could be taken using plot size or other important qualifying features of the studies. Other issues, such as what measures are appropriate surrogates for productivity, are scientific rather than statistical matters, and we do not comment on those here.

Based on this discussion, it is obvious that different meta-analysis methods are applicable in different situations. The informed meta-analyst, then, has at his or her disposal a progression of increasingly comprehensive models and methods. It is incumbent not to stop at simple vote counting, but equally not to use more quantitative methods without suitable data or satisfaction of assumptions. We therefore recommend that "best statistical practice" is included as part of the evolving "best practice" of meta-analysis. This embeds an exploratory phase that allows for qualitative discussion of comparable estimates from different studies (which may include vote counting) followed by an inferential phase that allows for different types of statistical modeling and analysis, based on what is supported by the data and by scientific understanding. In both phases, the principles that are now recognized as underpinning all of the "best practice" components of meta-analysis will be expected; that is, the methods that are adopted must be stated clearly, underlying assumptions defended and caveats about the limitations of the methods acknowledged.

Ecological relationships are invariably complex, so it is not unexpected that meta-analysis is difficult in this discipline area; however, this is also why meta-analysis can be a powerful tool for providing an overall, informed opinion about the collected body of literature. Similar issues to the ones in this Forum have been discussed in the context of meta-analysis and systematic review in other disciplines (e.g., see reviews by Mullen and Ramírez 2006, Quintana and Minami 2006). As is evident from recent literature, the identification of problems with the application in ecology of existing statistical techniques for meta-analysis motivates the development of new techniques, which in turn motivates increased, informed adoption of these techniques in the scientific community. Thus instead of merely denigrating the current state of the art, progress is more likely to result from working to develop models and methods for systematic review and quantitative research synthesis that are practical, applicable, valid, and robust from both ecological and statistical perspectives.

## Literature Cited

Arends, L. R. 2006. Multivariate meta-analysis: modeling the heterogeneity mixing apples and oranges; dangerous or delicious? Dissertation. Erasmus University, Rotterdam, The Netherlands.

Bagnardi, V., A. Zambon, P. Quatto, and G. Corrao. 2004. Flexible meta-regression functions for modeling aggregate dose-response data, with an application to alcohol and mortality. American Journal of Epidemiology 159:1077–1086.

Becker, B. J., and M. J. Wu. 2007. The synthesis of regression slopes in meta-analysis. Statistical Science 22:414–429.

Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. 2009. Introduction to meta-analysis. John Wiley and Sons, Chichester, West Sussex, UK.

Cooper, H., L. V. Hedges, and J. C. Valentine, editors. 2009. The handbook of research synthesis and meta-analysis. Second edition. Russell Sage Foundation, New York, New York, USA.

Fahrbach, K. R. 2001. An investigation of methods for mixed-model meta-analysis in the presence of missing data. Dissertation. Michigan State University, East Lansing, Michigan, USA.

Gillman, L. N., and S. D. Wright. 2006. The influence of productivity on the species richness of plants: a critical assessment. Ecology 87:1234–1243.

Gillman, L. N., and S. D. Wright. 2010. Mega mistakes in meta-analyses: devil in the detail. Ecology 91:2550–2552.

Goldstein, H. 2003. Multilevel statistical methods. Third edition. Edward Arnold, London, UK.

Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. Ecology 80:1142–1149.

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

Hedges, L. V., and I. Olkin. 1985. Statistical methods for meta-analysis. Academic Press, San Diego, California, USA.

Helser, T. E., and H.-L. Lai. 2004. A Bayesian hierarchical meta-analysis of fish growth with an example for North American largemouth bass, *Micropterus salmoides*. Ecological Modelling 178:399–416.

Hillebrand, H., and B. J. Cardinale. 2010. A critique for meta-analyses and the productivity–diversity relationship. Ecology 91:2545–2549.

Ioannidis, J. P., J. C. Cappelleri, and J. Lau. 1998. Meta-analyses and large randomized, controlled trials. New England Journal of Medicine 338:59.

Ives, A. R., and S. R. Carpenter. 2007. Stability and diversity of ecosystems. Science 317:58–62.

Jones, A. T., W. N. Venables, I. B. Dry, and J. T. Wiskich. 1994. Random effects and variances: a synthesis of nonlinear regression analyses of mitochondrial electron transport. Biometrika 81:219–235.

Laanisto, L., P. Urbas, and M. Pärtel. 2008. Why does the unimodal species richness–productivity relationship not apply to woody species: a lack of clonality or a legacy of tropical evolutionary history? Global Ecology and Biogeography 17:320–326.

Lajeunesse, M., and C. Schmid. *In press*. Recovering missing or partial data from studies: a survey of conversions and imputations for meta-analysis. *In* J. Koricheva, J. Gurevitch, and K. Mengersen, editors. Handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton, New Jersey, USA.

Lau, J., E. M. Antman, J. Jimenez-Silva, B. Kupelnick, F. Mosteller, and T. C. Chalmers. 1992. Cumulative meta-analysis of therapeutic trials for myocardial infarction. New England Journal of Medicine 327:248–254.

Lau, J., H. Rothstein, and G. A. Stewart. *In press*. Meta-analysis in medicine, social sciences, and ecology and evolution. *In* J. Koricheva, J. Gurevitch, and K. Mengersen, editors. Handbook of meta-analysis in ecology and evolution. Princeton University Press, Princeton, New Jersey, USA.

LeLorier, J., G. Gregoire, A. Benhaddad, J. Lapierrer, and F. Derderian. 1997. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. New England Journal of Medicine 337:536–543.

Lipsey, M. W., and D. B. Wilson. 1993. The efficacy of psychological, educational and behavioral treatment: confirmation from meta-analysis. American Psychologist 48:1181–1209.

Mitchell-Olds, T., and R. G. Shaw. 1987. Regression analysis of natural selection: statistical influence and biological interpretation. Evolution 41:1149–1161.

Mittelbach, G. G., S. M. Scheiner, and C. F. Steiner. 2003. What is the observed relationship between species richness and productivity? Reply. Ecology 84:3390–3395.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and

L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381–2396.

Mullen, P. D., and G. Ramírez. 2006. The promise and pitfalls of systematic reviews. Annual Revue of Public Health 27:81–102.

Pärtel, M., L. Laanisto, and M. Zobel. 2007. Contrasting plant productivity–diversity relationships across latitude: the role of evolutionary history. Ecology 88:1091–1097.

Paul, P. A., P. E. Lipps, and L. V. Madden. 2006. Meta-analysis of regression coefficients for the relationship between fusarium head blight and deoxynivalenol content of wheat. Ecology and Epidemiology 96:951–961.

Pigott, T. D. 2009. Handling missing data. Pages 399–416 *in* H. Cooper, L. V. Hedges, and J. C. Valentine, editors. The handbook of research synthesis and meta-analysis. Second edition. Russell Sage Foundation, New York, New York, USA.

Quintana, S. M., and T. Minami. 2006. Guidelines for meta-analyses of counseling psychology research. Counseling Psychologist 34:839–877.

Raudenbush, S. W., and A. S. Bryk. 2002. Hierarchical linear models: applications and data analysis methods. Second edition. Sage Publications, Thousand Oaks, California, USA.

Riley, R. D., K. R. Abrams, A. J. Sutton, P. C. Lambert, and J. R. Thompson. 2007. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Medical Research Methodology 7(3).

Riley, R. D., J. R. Thompson, and K. R. Abrams. 2008. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. Biostatistics 9:172–186.

Sharpe, D. 1997. Of apples and oranges, file drawers and garbage: why validity issues in meta-analysis will not go away. Clinical Psychology Review 17:881–901.

Simpson, G. G., A. Roe, and R. C. Lewontin. 1960. Quantitative zoology. Revised edition. Harcourt, Brace and Company, New York, New York, USA.

Slavin, R. E. 1986. Best evidence synthesis: an alternative to meta-analytic and traditional reviews. Educational Research 15:5–11.

Slavin, R. E. 1995. Best evidence synthesis: an intelligent alternative to meta-analysis. Journal of Clinical Epidemiology 48:9–18.

Sutton, A. J., K. R. Abrams, D. R. Jones, and F. Song. 2000. Methods for meta-analysis in Medical Research. John Wiley and Sons, Chichester, UK.

Thompson, N., and R. H. Hobbs. 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecological Applications 16:5–19.

Tweedie, R. L., and K. L. Mengersen. 1995. Meta-analysis approaches to dose–response relationships in studies of lung cancer and passive smoking. Statistical Medicine 14:545–569.

Tylianakis, J. M., R. K. Didham, J. Bascompte, and D. A. Wardle. 2008. Global change and species interactions in terrestrial ecosystems. Ecology Letters 11:1351–1363.

van Houwelingen, H. C., L. R. Arends, and T. Stijnen. 2002. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine 21:589–624.

West, B., K. B. Welch, and A. T. Galecki. 2007. Linear mixed models: a practical guide using statistical software. Chapman and Hall/CRC Press, New York, New York, USA.

Whittaker, R. J. 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91:2522–2533.

Whittaker, R. J., and E. Heegaard. 2003. What is the observed relationship between species richness and productivity? Comment. Ecology 84:3384–3390.

Witman, J. D., M. Cusson, P. Archambault, A. J. Pershing, and N. Mieszkowska. 2008. The relation between productivity and species diversity in temperate-arctic marine ecosystems. Ecology 89(Supplement):S66–S80.

Wolpert, R., and K. Mengersen. 2004. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. Statistical Science 19:450–471.

*Ecology*, 91(9), 2010, pp. 2561–2564
© 2010 by the Ecological Society of America

# 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.

## 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,

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

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 non-random 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.

*Ecology*, 91(9), 2010, pp. 2565–2567
© 2010 by the Ecological Society of America

# The productivity–diversity relationship: varying aims and approaches

Meelis Pärtel,[1] Kristjan Zobel, Lauri Laanisto, Robert Szava-Kovats, and Martin Zobel

*Institute of Ecology and Earth Sciences, University of Tartu, Lai 40, Tartu 51005 Estonia*

The productivity–diversity relationship is an important and heavily debated issue in ecology. A major advancement in exploring this relationship has been the analyses of large numbers of case-studies (Mittelbach et al. 2001, Gillman and Wright 2006). These works have shown that the relationship between productivity and diversity varies, and that a single "general" relationship does not exist. We investigated the productivity–diversity relationship with respect to the species pool concept (Pärtel et al. 2007), which postulates that more species are expected to evolve in conditions (ecosystems) that have been historically more common. Tropical humid ecosystems have been relatively more productive during the past hundreds of millions of years, whereas productivity in temperate ecosystems has been limited by low temperatures and repeated glaciations. We found that positive productivity–diversity relationships are more common in the tropics, where the species pool of productive habitats is large, whereas declines in species richness at high productivities is more common in temperate regions, since species pools in high productivity ecosystems are likely to be relatively small. Similarly, we showed that positive productivity–diversity relationship can be common even in temperate regions, if only woody species diversity is considered (Laanisto et al. 2008). This finding might reflect evolutionary history: most temperate woody species originate from tropical lineages and exhibit "tropical" patterns due to niche conservatism.

Robert J. Whittaker (2010, hereafter RJW) criticizes our above-mentioned studies, claiming we were too liberal in selecting case studies. RJW's argument relies on the assumption that the seven conditions for case studies he describes as "reasonable and *necessary* criteria," are always valid. We argue that different research aims require a priori criteria to be defined. Demanding implementation of criteria as hindsight criticism of existing studies can easily lead to paradoxes (e.g., richness rather than diversity should be used for *diversity* relationships). Our approach in selecting case studies for analysis was straightforward: a case study must reveal a suitable habitat-productivity–plant-diversity relationship. In addition, although having diverse studies increases the amount of noise in the data, false rejections of the null hypotheses does not increase unless there is a systematic bias induced to productivity or diversity. In the following, we address each criterion suggested by RJW, detailed responses to criticism of particular papers used in our study are found in the Appendix.

*Species richness is the only acceptable measure of diversity.*—We found diversity more appropriate because diversity encompasses both richness and several other diversity metrics. In addition, different diversity measures are often strongly correlated. Nor do we see any complication if only a prominent subset of total plant diversity is used. Taxonomic and functional limits are always artificial. If we define plants as autophototrophic organisms, then in order for the data to be "complete," we would be obliged to include bryophytes, algae, and cyanobacteria as well. Moreover, plant diversity can only be estimated (some species are dormant, some possible misidentified during sampling). In any case, to study the productivity–diversity relationship as we did, this criterion is simply not applicable.

*Plot size must be constant.*—Indeed, diversity is traditionally estimated per fixed area. However, because plant individuals vary in size, it might be more appropriate to measure diversity per fixed number of ramets (Oksanen 1996). The general productivity–diversity relationship, however, does not necessarily change when diversity is measured from both fixed area and fixed number of ramets (Zobel and Liira 1997, Liira and Zobel 2000). As a rule, we preferred diversity from fixed area.

*Adequate surrogate for productivity.*—Productivity is rarely measured directly as the rate of carbon flux through plants or animals (Cardinale et al. 2009), rather different proxies are used (biomass, soil resources, precipitation, and so on). However, no effect of productivity proxy has been found on the pattern of the productivity–diversity relationship (Groner and Novoplansky 2003, Pärtel et al. 2007).

*Data distribution.*—This is a fine suggestion, but not examined by RJW in depth.

*Confounding factors.*—This is definitely a necessary criterion that we adhered to. RJW rejects many case studies that include grazing or wildfires. Nevertheless, all ecosystems feature characteristic disturbance regimes

Forum

Manuscript received 9 June 2009; revised 13 July 2009; accepted 29 July 2009. Corresponding Editor: D. R. Strong. For reprints of this Forum, see footnote 1, p. 2534.

[1] E-mail: meelis.partel@ut.ee

and have developed under these conditions. Thus, we cannot support the notion to omit a case study if other environmental or disturbance gradients are mentioned in a paper, unless there is clear evidence of confounding effects.

*Number of data points*.—Setting limit $n \geq 10$ has absolutely no statistical justification. We rely on statistical testing that either succeeds or fails to reject the null hypothesis.

*A single data set is included only once*.—This is a statistically sound criterion. We agree with RJW that sometimes it is difficult to trace data sources (see the Appendix). RJW expands on this point with respect to scale. How should one treat a single system studied at different scales? We compromised: if the result was the same at different scales, we reported this only once, and if the result varied across scales, we used two data points since we cannot define a priori a "correct" scale.

RJW reports major discrepancies between our results and his (his Table 2). This is based on the grouping of productivity–diversity relationships. RJW accepts the grouping of productivity–diversity relationships into five patterns (positive, negative, unimodal, U-shaped, no relationship) and criticizes our classification in which we merged negative with unimodal and U-shaped with no relationships. Regardless of how we estimate productivity, at zero productivity there is de facto zero diversity. Ideally, all trends between productivity and diversity should originate at the zero-zero point. For practical reasons, however, simpler models can be calculated to accommodate the particular range of productivity found in a data set. Nevertheless, the merging of unimodal and negative relationships makes ecological sense, since both patterns show that diversity declines with increasing productivity. Although RJW is correct that U-shaped and unimodal patterns are mathematically equal, these patterns are not equal ecologically. We know of no applicable ecological explanations for U-shaped patterns. We assert that—dependant on the aim of the study—different classifications are both acceptable and valid.

Pertinent to the previous point, we must indicate inaccuracies in RJW's Table 3. The table erroneously contains zeros for our paper under the columns "negative" and "U-shaped." These must be designated as "not applicable," since we did not use these classes. In addition, zero papers are marked as "inadmissible." This is correct for the subset considered by RJW, but not for similarity calculations presented in Table 2. Although we used references from Mittelbach et al. (2001), we scrutinized all papers ourselves and searched for additional case-studies. There were in fact hundreds of papers we classified as "inadmissible"; we simply saw no need to mention the fact. Therefore, similarities reported by RJW in his Table 2 should also consider the multitude of papers that failed our acceptance criteria. Most of these studies would likely have been classified as inadmissible by RJW as well, and the similarity between his and our results would be extremely high!

Nevertheless, we can still explore similarities among the different approaches using the comparable part of RJW's Table 3. What is the expected proportion of positive productivity–diversity relationships compared to relationships where at least some productivity range diversity declines with increasing productivity (unimodal or negative relationship)? RJW presents interpretations from different sources (his Table 3). This table is based largely on our appendix in Pärtel et al. (2007) and these data sets have been interpreted by both their original authors and RJW. Of the subset selected by RJW (his Table 3) we interpreted 15 positive and 34 unimodal (including negative) relationships. Originally the authors of these studies distinguished 8 positive relationships and 15 unimodal (including negative) relationships. The lower total number is due to the fact that the data were often provided in a case study with different purposes. RJW casts aside most of the studies we used, and of the remaining perceives 5 positive and 10 unimodal (including negative) relationships. If we compare these numbers, we find an unquestionable consensus in proportions (e.g., Fisher exact test for a 3 × 2 table gives $P = 0.951$). Accordingly, there is a truly "mega-result" for this subset of case studies on productivity–diversity relationships: all interpretations report relationships 1/3 positive and 2/3 unimodal (including negative) relationships!

We acknowledge that the linkage between species-area relationship and productivity is a very interesting avenue to study. Positive productivity–diversity relationships are often found at larger spatial extent (Gillman and Wright 2006). In contrast, Mittelbach et al. (2001) have addressed the question whether productivity–diversity relationship might depend on grain (sample plot size) but found no differences. Thus, we conclude that unless clear patterns with grain, focus, and extent become evident, studies encompassing multiple scales remain legitimate. We excluded, however, continental- and global-scale studies since these varied much in their evolutionary background.

We certainly cannot agree with the idea that plot size for plant diversity studies should be canonized. We recognize the need for standards for phytosociological classification of vegetation (Chytry and Otypkova 2003). If ecological community would limit biodiversity data to $>16$ m$^2$ in grasslands and $>200$ m$^2$ in woodlands, as suggested by RJW, we would have to discount most biodiversity studies ever published! The aim of phytosociology is to describe species composition and plot size suggestions were never meant for diversity studies (M. Chytry, *personal communication*). We are confident that biodiversity relationships with other ecological gradients are equally interesting, starting from point estimates and ending with the global scale.

To sum up, how similar should the case studies be in order to draw sound generalizations? The answer depends on the question: if we wish to address the local mechanisms underlying the patterns, we might limit ourselves to data sets with uniform measurements from a single vegetation type or even a location. Conversely, if we aim to address global diversity patterns, we need to tackle multiple approaches and methods. The more we extend the scale of observation or the scope of study, the more diverse the case studies will be. This is no reason to capitulate, because statistics are meant to deal with indefinite observations! Therefore we remain confident that our "lower similarity" conception is an adequate approach considering our aims.

What suggestions can we offer? We certainly agree with RJW of the need for new rigorous experimental and field studies on productivity–diversity relationships. Field studies are unfortunately very difficult to perform if we want to repeat the study in several regions across the world. We are, however, far less enthusiastic about RJW's other suggestions. There are already many narrative reviews and such compilations can contain even more dangerous "dragons" than seen by RJW in our works. In addition, we are skeptical whether such reviews can address the questions we studied, e.g., how ecological relationships vary across latitude. The idea of RJW to use the best evidence synthesis is novel to ecology. Not a single published article can be found from ISI Web of Science when searching for the "best evidence synthesis" within ecological periodicals. Therefore, it seems premature to call for an end to the old method before the new one has been established, and we hope that there is a place for multiple approaches in our friendly ecological community.

## Acknowledgments

This research was supported by the European Union through the European Regional Development Fund (Center of Excellence FIBIR) and by the Estonian Science Foundation (Grants 6614, 7576, and 7016).

## Literature Cited

Cardinale, B. J., H. Hillebrand, W. S. Harpole, K. Gross, and R. Ptacnik. 2009. Separating the influence of resource "availability" from resource "imbalance" on productivity–diversity relationships. Ecology Letters 12:475–487.

Chytry, M., and Z. Otypkova. 2003. Plot sizes used for phytosociological sampling of European vegetation. Journal of Vegetation Science 14:563–570.

Gillman, L. N., and S. D. Wright. 2006. The influence of productivity on the species richness of plants: a critical assessment. Ecology 87:1234–1243.

Groner, E., and A. Novoplansky. 2003. Reconsidering diversity–productivity relationships: directness of productivity estimates matters. Ecology Letters 6:695–699.

Laanisto, L., P. Urbas, and M. Pärtel. 2008. Why does the unimodal species richness–productivity relationship not apply to woody species: a lack of clonality or a legacy of tropical evolutionary history? Global Ecology and Biogeography 17:320–326.

Liira, J., and K. Zobel. 2000. The species richness–biomass relationship in herbaceous plant communities: what difference does the incorporation of root biomass data make? Oikos 91:109–114.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381–2396.

Oksanen, J. 1996. Is the humped relationship between species richness and biomass an artefact due to plot size? Journal of Ecology 84:293–295.

Pärtel, M., L. Laanisto, and M. Zobel. 2007. Contrasting plant productivity–diversity relationships across latitude: the role of evolutionary history. Ecology 88:1091–1097.

Whittaker, R. J. 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91:2522–2533.

Zobel, K., and J. Liira. 1997. A scale-independent approach to the richness vs biomass relationship in ground-layer plant communities. Oikos 80:325–332.

## APPENDIX

Detailed reply to criticism from Robert J. Whittaker, provided in his Appendix A of "Meta-analyses and mega-mistakes: calling time on meta-analyses of the species richness–productivity relationship" (*Ecological Archives* E091-186-A1).

*Ecology*, 91(9), 2010, pp. 2568–2571

# In the dragon's den: a response to the meta-analysis forum contributions

Robert J. Whittaker[1]

*Biodiversity Research Group, Oxford University Centre for the Environment, South Parks Road, Oxford OX1 3QY United Kingdom* and *Centre for Macroecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark*

Secondary analysis of previously published data has a long tradition in ecological science and is widely and successfully practiced as a means of efficiently addressing new questions and hypotheses. Meta-analysis is, in essence, the class of such analyses in which the findings of multiple primary studies are subject to further statistical analysis of emergent outcomes, and is a more recent practice within ecology. I recognize that this is a loose definition of meta-analysis (Ellison 2010, Gurevitch and Mengersen 2010) but continue to refer to the studies I critique using this common broader usage. Owing to the apparent power of such synthetic analyses, meta-analysis papers can be highly influential (Mittelbach 2010). This forum, together with other recent critical assessments (e.g., Englund et al. 1999, Gates 2002), demonstrates that there are good reasons to call for great care, improved rigor and transparency in the use of "meta-analysis" tools in ecology. However, in the article that initiated this forum exchange (Whittaker 2010), all the specific criticisms I made were restricted to recent meta-analyses of just one problem, which concerns the form of the species richness–productivity relationship (SRPR) in plants. In this brief response to the seven other contributions, I retain this focus while aiming to resolve several misconstructions of points made in my paper, and to comment on a few key points of disagreement regarding analyses of the SRPR.

*Use of proxies.*—First, there have been relatively few studies that have specifically set out to gather data to determine the form of the SRPR and so, in order to increase the power of analysis and refine the questions asked, those undertaking meta-analyses have sought other data sets that were initially gathered for different purposes. There are many published papers providing diversity data, but few that provide direct measurements of productivity, which is a difficult property to estimate accurately. Hence the reliance in Mittelbach et al. (2001), Pärtel et al. (2007), and Laanisto et al. (2008) on the use of proxies such as rainfall, vegetation height, biomass, etc., in order to generate surrogate productivity data for their analyses of the SRPR. Unfortunately, nonlinearities in relationships between actual productivity and the productivity proxies used in these analyses have the potential to result in misclassification of the form of the SRPR (see: Whittaker and Heegaard 2003, Gillman and Wright 2006, 2010, Huston and Wolverton 2009). By reference to details drawn from the original source papers and the wider literature, I have argued that this seriously undermines the analyses (Whittaker 2010: Appendix A). This was only one of a number of reasons leading me to stress the necessity of screening data sets for fitness-for-purpose prior to inclusion in analysis.

*Criteria for selecting data sets.*—Other contributors to the forum regard my criteria for including a data set in an SRPR meta-analysis as too limiting. For instance, Lajeunesse (2010) argues that "Erroneous elimination of a prohibitive number of studies is not a solution to handling variation due to study 'quality'..." Instead, we should "... gather all studies relevant to the conceptual topic under study, and then empirically test whether these differences ... actually influence research outcomes." However, the published SRPR meta-analyses demonstrate that different authors have adopted very different views of the "relevance" of an original study. Mittelbach et al. (2001) developed and reported a search strategy based on key words, e.g., a paper would have been screened if it had "species richness" in the key words but then rejected if it turned out there were no data *they* felt able to use as productivity proxies. They also eliminated studies of systems subject to severe anthropogenic disturbance, etc. By contrast, Pärtel et al. (2007) and Laanisto et al. (2008) did not reveal their criteria, and used many studies, that are not "relevant to the conceptual topic" and which in my view do not provide *suitable* data, free from confounding problems such as anthropogenic manipulation. It is thus of little practical help to say that we should use "all relevant studies" and then see if the (very many) factors identified as problematic have a statistical influence: the meta-analyst has first to decide and justify *which* are relevant (Gates 2002). Similarly, to provide a specific example, it

Manuscript received 10 September 2009; revised 19 November 2009; accepted 16 December 2009. Corresponding Editor: D. R Strong. For reprints of this Forum, see footnote 1, p. 2534.

[1] E-mail: robert.whittaker@ouce.ox.ac.uk

is no answer to the serious lack of standardization of sampling in Beadle's (1966) data set to say, as do Pärtel et al. (2010: Appendix A), that because Beadle saw fit to plot a regression line (actually it appears to be hand fitted) through a set of values, it is therefore safe to use for this new purpose. I therefore reiterate the view that a key initial step in meta-analysis should be to develop, articulate and apply a set of criteria for determining the studies that are to be included in the analysis.

I recognize that other ecologists may accept the need to have stated criteria while disagreeing with the particular set that I put forward for future use (e.g., see Hillebrand and Cardinale 2010). Here, I aim to clarify some of my choices regarding data set eligibility criteria. Criterion 1: I did not state that analysis of diversity indices are wrong, merely that alpha diversity indices provide different response variables, distinct from species richness, and that different response variables should be analyzed in separate (meta)analyses. Criterion 2: I may not have worded this clearly enough. My argument is that for a particular data set to be included in the meta-analysis, the plots reported *in that data set* should be of a fixed size (I suggested within ±10%, but with very small plots within ±5%). Holding plot size constant is necessary when sampling plant species richness because increasing the contiguous sample area from a small plot size to increasingly large plot sizes inevitably involves a stepped pattern of increased richness with area: failure to hold plot size constant within a data set means that area confounds analysis. I should emphasize that the criteria under discussion here are those I suggest for screening data for inclusion. In my article I also emphasized the need to organize the meta-analysis step with reference to scale of the study systems, but this is a separate step from screening individual studies for eligibility. Criterion 5 states that data sets involving other prominent confounding variables should be screened out, and I gave the examples of mowing, grazing, horticulture, or burning. An alternative to removing such studies is to examine whether the variable has explanatory power for the form of the SRPR. This approach has been adopted, for example, in examining the role of mesh size in meta-analyses of stream predation experiments (Englund et al. 1999). This may be tractable in systems in which there is a general consistency of approach and a limiting number of fairly obvious confounding factors. The difficulty presented in meta-analyses of the SRPR in plants, is that there appear to be so many potential confounding factors in the data sets gathered, that it becomes analytically intractable to deal with all of them at the formal meta-analysis step. Criterion 6 is the imposition of a minimum number of data points (within a particular study data set) for inclusion in these meta-analyses. Hillebrand and Cardinale (2010) argue that imposition of a 10-data-point minimum requirement is arbitrary and unnecessarily restrictive. Perhaps they are right, but the data sets in these analyses are noisy, productivity proxies are problematic, confounding variables are rarely entirely out of the equation, and inclusion of four- or five-point data sets in analyses testing between humped and linear fits seems risky in this context. This is why both Mittelbach et al. (2001) and Gillman and Wright (2006) used this criterion in their meta-analyses of the SRPR: I merely adopted their suggestion. Hillebrand and Cardinale sum up that my suitability criteria are "very arbitrary" but disappointingly do not provide their own alternative, less arbitrary set.

*The logic of collapsing categories*.—The first of the SRPR meta-analyses, by Mittelbach et al. (2001), set out to classify each SRPR as one of (1) positive linear, (2) humped, (3) negative linear, (4) U-shaped, and (5) unclassifiable, basing their decisions on standardized statistical procedures. In their analyses, Pärtel et al. (2007) collapse these categories. They assign U-shaped SRPR (which are very rare) to unclassifiable, on the grounds that they cannot see how to theorize a u-shaped relationship. They assign negative SRPR to humped SRPR on the basis that studies returning negative SRPR have probably not sampled environments of sufficiently low productivity to reveal the initial rising limb of what they theorize to be the real hump-shaped form. This sampling bias hypothesis is not a supportable generalization based on the source literature I have examined (Whittaker 2010: Appendix A). Moreover, if this logic is deemed acceptable, then why should we not convert positive linear relationships to humped relationships? Here the logic would be that systems showing positive linear relationships must merely have failed to sample high enough productivities to display the downwards part of the curve. This is as inherently plausible an argument as that concerning negative relationships, and, like that argument, may well apply in some cases (but in *which* and *how many* cases is unknowable). As these two arguments are logically equivalent, accepting one implies accepting the other, meaning that if statistical analysis reveals any one of forms 1, 2, or 3 (positive, humped, or negative), they should be (re-)classified as a humped SRPR; while other studies would be deemed to belong to the "no relationship" group. The humped SRPR then becomes general (the proposition Mittelbach et al. 2001 set out to test), but by proclamation rather than by statistical analysis. To pursue such arguments is to allow our beliefs about the likely true form of an unsampled portion of a relationship to hold sway over statistical analysis carried out within the empirical range of the study systems we have analyzed. This is unwarranted and, if undertaken, may easily be misunderstood by readers.

*Agreements and disagreements on the detail*.—As Gillman and Wright (2010) point out, we concur in most matters and there is a strong measure of agreement between our decisions on the form of the SRPR (but for differences, see Whittaker 2010: Appendix A case studies 10, 106/108, 131, 147, 151, 152, and 157). I thank them for pointing out my error in incorrectly transcribing

Mittelbach et al.'s classification of the study by Wheeler and Shaw (1991), although it remains the case that I differ from Gillman and Wright (2006) in regarding it as more likely a negative rather than U-shaped relationship. Bear in mind that my classification is based solely on reading the source paper and visual examination of the data set, not on new statistical analyses. In this instance, the data set includes a lot of scatter and has been variously regarded as negative by the original authors (and by me), humped (because negative is taken to mean humped) by Pärtel et al. (2007), and U-shaped by Mittelbach et al. (2001) and Gillman and Wright (2006). The limited consensus in this case is merely that the relationship is not positive.

Notwithstanding the concerns I have raised, Pärtel et al. (2010) repeat their claims to have demonstrated a tropical vs. temperate difference in the form of the productivity–diversity relationship, strongly implying that causation of this difference is related to species pool size. Here, indeed, be dragons. Their responses, especially as set out in their Appendix, provide revealing insights into the hitherto unstated criteria used by this team of authors and serve to illustrate that their data base cannot withstand forensic scrutiny. I find little scope for a more positive assessment of their findings in the light of this defense and recommend that any interested readers call up the source papers from online journal resources and archives, to evaluate how these data have been used in each meta-analysis.

*Whither meta-analyses of the SRPR*.—Those reviewing the first draft of this manuscript questioned whether it was productive to continue debating perceived flaws in the treatment of the SRPR. I sympathize with this perspective, but have invested in doing so because meta-analyses tend to carry influence and to become highly cited: they shape understanding and opinion disproportionately. So, for example, Oberle et al. (2009:6–7) comment that "… Recent work has shown that in herbaceous plant communities, clonal species may dominate high-productivity environments, increasing the prevalence of hump-shaped SRPRs, while this trait is less common among woody growth forms, resulting in more monotonic SRPRs …" In support of this statement they cite the paper by Laanisto et al. (2008), which itself is a reworking and extension of the Pärtel et al. (2007) data base. I submit that while the foregoing statement could be correct at some scale of analysis, no such inference can reliably be based upon this particular source (see Oberle et al. [2009] for further discussion).

In some respects, I think we are seeing a classic trade-off here. In regular empirical papers, the reader gets to see the details of the sampling regime, study site, and key assumptions and can readily assess the strength of the inferences drawn. Such studies have value, but on their own provide singular cases that may not be representative. Meta-analyses (including quantitative data synthesis papers that are not technically meta-analyses), undoubtedly have greater agency (influence) than most primary data papers, but the properties of the underlying data are less easy for the reader to detect and scrutinize. This means, as other forum contributors argue, that it is doubly important that all key assumptions and analytical steps are clearly stated and that the meta-data are treated with great care by the meta-analysts (Gates 2002). The challenges involved for those involved in meta-analysis preparation and review are thus—like the influence such papers may have—disproportionate.

*The scale issue*.—I concur with a great deal of Mittelbach's (2010) thoughtful essay, although we continue to differ in our perspectives regarding scale, wherein I place relatively greater emphasis on the focal scale of analysis as an organizing principal in meta-analysis. On this issue, Mittelbach (2010) cites a specific study based on two data sets, demonstrating scale-invariance in the shape of the SRPR over a focal scale range of 10 m$^2$ to 200 m$^2$. However, as he recognizes, we cannot know if that scale-invariance would continue outside this empirical range, or for other systems. Other studies discussed by Whittaker (2010) do show (focal-)scale dependency. Notwithstanding our differences of perspective, I concur with Mittelbach's comments on Chase and Leibold (2002) as, in this instance, change in the form of the SRPR did not arise from changing plot sizes but rather from aggregating sites. This indicates that such changes in form can arise in studies of varying data structure, a common component being that as focal scale changes different diversity components are implicated.

Notwithstanding the concerns I have expressed about many of the case studies (Whittaker 2010: Appendix A), there appear to be sufficient recent empirical studies of robust design, to allow us to conclude that for a particular place and study system extent, the form of the SRPR can and frequently does change from linear to unimodal or vice versa with changing focal scale of analysis (Whittaker 2010). This means, I suggest, that we cannot view *any* study based on *a single focal scale* of analysis as adequately characterizing the general form of the relationship for that *place*, *system*, or *extent*. At finer or coarser focal scales it is quite likely that the system will have a different form of SRPR. Thus, all other issues aside, we cannot yet make any claim as to (e.g.) geographical differences in the form of the SRPR, without controlling in analysis for focal scale used. I suspect that this is a problem that has a wider relevance than yet realized in the quest for understanding geographical patterns of diversity.

Finally, I would like to make two points of clarification of the section *A few tasters* in Whittaker (2010), arising from correspondence following acceptance. First, I am grateful to L. N. Gillman and S. D. Wright for pointing out that Mittelbach et al. (2001) in fact classified the Wheeler and Shaw (1991) data set as U-shaped; hence, the reported relationships should read Wheeler and Shaw negative; GW2006 and M2001 U-shaped; P2007 humped. The inclusion of a humped relationship for M2001 in Appendix A (and Table A1) of Whittaker (2010) is thus in error. Second, regarding the Wardle et al. (1997) study, the comment about island area variation warrants further exposition; although the Shannon-Weiner diversity index data were collected from standard-sized plots, these plots are derived from islands of strongly contrasting size, and within the source paper it is demonstrated that island area is a strong determinant of environmental and ecosystem (including long-term succeessional) dynamics, thus confounding the interpretation of causal relationships.

## Literature Cited

Beadle, N. C. W. 1966. Soil phosphate and its role in molding segments of the Australian flora and vegetation, with special reference to xeromorphy and sclerophylly. Ecology 47:992–1007.

Chase, J. M., and M. A. Leibold. 2002. Spatial scale dictates the productivity–biodiversity relationship. Nature 416:427–430.

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

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

Gates, S. 2002. Review of methodology of quantitative reviews using meta-analysis in ecology. Journal of Animal Ecology 71:547–557.

Gillman, L. N., and S. D. Wright. 2006. The influence of productivity on the species richness of plants: a critical assessment. Ecology 87:1234–1243.

Gillman, L. N., and S. D. Wright. 2010. Mega mistakes in meta-analyses: devil in the detail. Ecology 91:2550–2552.

Gurevitch, J., and K. Mengersen. 2010. A statistical view of synthesizing patterns of species richness along productivity gradients: devils, forests, and trees. Ecology 91:2553–2560.

Hillebrand, H., and B. J. Cardinale. 2010. A critique for meta-analyses and the productivity–diversity relationship. Ecology 91:2545–2549.

Huston, M. A., and S. Wolverton. 2009. The global distribution of net primary production: resolving the paradox. Ecological Monographs 79:343–377.

Laanisto, L., P. Urbas, and M. Pärtel. 2008. Why does the unimodal species richness–productivity relationship not apply to woody species: a lack of clonality or a legacy of tropical evolutionary history? Global Ecology and Biogeography 17:320–326.

Lajeunesse, M. J. 2010. Achieving synthesis with meta-analysis by combining and comparing all available studies. Ecology 91:2561–2564.

Mittelbach, G. G. 2010. Understanding species richness–productivity relationships: the importance of meta-analyses. Ecology 91:2540–2544.

Mittelbach, G. G., C. F. Steiner, S. M. Scheiner, K. L. Gross, H. L. Reynolds, R. B. Waide, M. R. Willig, S. I. Dodson, and L. Gough. 2001. What is the observed relationship between species richness and productivity? Ecology 82:2381–2396.

Oberle, B., J. B. Grace, and J. M. Chase. 2009. Beneath the veil: plant growth form influences the strength of species richness–productivity relationships in forests. Global Ecology and Biogeography 18:416–425.

Pärtel, M., L. Laanisto, and M. Zobel. 2007. Contrasting plant productivity–diversity relationships across latitude: the role of evolutionary history. Ecology 88:1091–1097.

Pärtel, M., K. Zobel, L. Laanisto, R. Szava-Kovats, and M. Zobel. 2010. The productivity–diversity relationship: varying aims and approaches. Ecology 91:2565–2567.

Wheeler, B. D., and S. C. Shaw. 1991. Above-ground crop mass and species richness of the principal types of herbaceous rich-fen vegetation of lowland England and Wales. Journal of Ecology 79:285–301.

Whittaker, R. J. 2010. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. Ecology 91:2522–2533.

Whittaker, R. J., and E. Heegaard. 2003. What is the observed relationship between species richness and productivity? Comment. Ecology 84:3384–3390.