Alkistis Elliott-Graves
a.elliott-graves@uni-bielefeld.de

Re-evaluating Heterogeneity in Evidence Synthesis

1. Introduction

In many disciplines such as Ecology, Evolutionary Biology and Medicine, there are often multiple individual studies addressing each research question. These are referred to as 'primary research' and are usually observational or experimental studies that directly investigate a question or phenomenon, by observing connections between causes and effects or actively manipulating systems to bring about certain effects. However, primary studies usually differ in a number of ways, for example, whether they are observational or experimental, whether the experiment is in a laboratory or in the field, which variables are being measured, the population and the size of the population being studied, and so on, so replicating the results of a study is rare (Hardwicke et al., 2020; Koricheva et al., 2013; Marsden et al., 2018; Mueller-Langer et al., 2019). Moreover, these differences often make comparing the results of the primary studies difficult. Scientists are increasingly using the 'evidence synthesis toolkit', which primarily consists of *systematic review* and *meta-analysis* to integrate and synthesise results from individual studies, so as to provide general answers to the original research questions, as well as information about their generality (Gurevitch et al. 2018, Koricheva et al. 2014, Stegenga 2011).

There has been an explosion of evidence synthesis in the literature, with many disciplines seeing an exponential increase in systematic review and meta-analysis papers between the 1990s and the 2010s (see for example Cadotte et al., 2012; Chen & Jhanji, 2012; Fontelo & Liu, 2018; Taylor & Munafò, 2016). Perhaps unsurprisingly, the reception of evidence synthesis has progressed along the typical trajectory of novel methodologies, with an initial period of hype, followed by a wave of skepticism about its role and usefulness in the greater context of scientific practice. Following some savage critiques of particularly problematic instances of meta-analysis (Ioannidis, 2005, 2016), the excitement surrounding the methodology dampened. Established scholars in various fields began to take on evidence synthesis, warning scholars not to rely on these methods as they would not only fail at their intended goals (e.g. of minimizing bias and helping to overcome the replication crisis) but could actively lower the overall quality of research in a field. The two main critiques are that

evidence synthesis perpetuates existing biases and introduces new types of bias, and that (Ioannidis, 2005, 2016; Romero, 2016; Stegenga, 2011; Watkins et al., 2021), and that evidence synthesis is often misleading due to the *heterogeneity* of primary research, i.e. that primary studies are too diverse to allow for meaningful comparisons or synthesis, so any synthesis will necessarily be flawed (Carpenter, 2020; Ioannidis et al., 2007; Spake et al., 2022; Whittaker, 2010).

Until recently, philosophers have remained largely absent from the furore surrounding evidence synthesis, with a few notable exceptions (Berchialla et al., 2020; Bruner & Holman, 2019; Fletcher, 2022; Holman, 2019; Jukola, 2017; Kovaka, 2022; Maziarz, 2022; Stegenga, 2011). Moreover, this philosophical coverage is quite patchy, as most papers focus only on one aspect of evidence synthesis (*meta-analysis*) and only within the context of *medicine* (Berchialla et al., 2020; Bruner & Holman, 2019; Fletcher, 2022; Holman, 2019; Jukola, 2017, 2017; Maziarz, 2022). In addition, philosophical accounts tend to side with the skeptics, emphasizing the misuses of meta-analysis rather than its potential value (Jukola, 2017; Maziarz, 2022; Romero, 2016; Stegenga, 2011). Finally, there have been no thorough philosophical examinations of the issue of heterogeneity in evidence synthesis, which take into account evidence synthesis in biology as well as medicine and the social sciences.

The aim of this paper is to re-examine the issue of heterogeneity in evidence synthesis. While the original critiques of heterogeneity highlight some valid points, I will argue that these points are most relevant when the main goal is to *generate causal confidence*, which usually occurs in the field of medicine. However, evidence synthesis can be used for different purposes, such as *arbitrating between contradictory results* and *exploring the scope of generalisations*, as is often the case in evolutionary biology, ecology and conservation. In cases like these, heterogeneity is less problematic than it has hitherto been portrayed, and can actually be positive, in the sense that it can provide useful information and even, on occasion, yield novel insights. I begin by providing a short overview of the process of evidence synthesis, as a context for understanding where heterogeneity comes in and how it is treated (section 2.1) followed by an outline of the critiques against heterogeneity in evidence synthesis (section 2.2). I then show that evidence synthesis is used for different purposes, which usually align with different disciplines (e.g. medicine vs biology) (section 3). In section 4, I explain when and why heterogeneity is genuinely problematic, and in section 5 contrast these cases with some where heterogeneity is not detrimental, and even, on occasion valuable. Section 6 provides some concluding remarks.

2. Heterogeneity and its Measurement

The worry with heterogeneity in evidence synthesis is that too many differences between primary studies renders them not comparable to each other (Carpenter, 2020; Ioannidis et al., 2007; Spake et al., 2022; Whittaker, 2010). Primary studies differ in many ways, in terms of their inputs (such as experimental setup, type of intervention, length of treatment, phenomenon/species/taxon being studied) and outputs (such as effect size, magnitude/direction of effect, how the effect is measured/presented). If studies are *too* different, then attempts at synthesis can be difficult, misleading or even completely meaningless. Before we can delve into the argument itself, however, it is important to understand the process of evidence synthesis, and when, how, and where heterogeneity manifests in this process.

2.1. Heterogeneity in context

The evidence synthesis toolkit consists of two main tools: *systematic review* and *meta-analysis*. These tools are not entirely independent from each other, in the sense that meta-analyses include the steps that constitute systematic reviews, yet systematic reviews are a legitimate stand-alone tool for evidence synthesis (see Table 1). A systematic review has three main components: the formulation of the research question, the search of the literature for original research on the topic and the decision of which of the available literature is relevant for the research question and will be included in the review (Foo et al., 2021). We can think of these components as steps, though in practice, they do not proceed iteratively, as each step is revised and refined multiple times throughout the review (Booth et al., 2012). Important/challenging issues in these steps are:

(i) Ensuring that the literature search is as broad as possible, by including multiple search engines
(ii) Ensuring that the literature search is efficient and transparent, by documenting the exact search terms used and
(iii) Making (documented) decisions about whether or not/and to what extent to include additional literature (e.g. grey literature, literature in other languages etc.).

All these decisions will affect the overall amount of heterogeneity in the sample of papers. The larger the basic pool of papers is, the more heterogeneous they are likely to be, as a larger pool of papers increases the likelihood that there will be differences between inputs (experimental setup, species studied, dosage etc). Any differences in inputs are likely to result in differences of outputs (effect size, magnitude/direction of effect etc.).

At this point, researchers can write up their findings in the form of a systematic review, where they discuss the papers that have not been discarded, or continue to conduct a full-blown

meta-analysis. The term 'meta-analysis' was coined in 1976 to describe research in the medical and social sciences that combined data from multiple studies to determine an 'overall effect' (Nakagawa et al., 2017). First, researchers extract the data from the primary studies and calculate each study's *effect size*, i.e. a statistical parameter that can be used to compare the results of different studies in which a common effect of interest has been measured (Koricheva et al., 2013). In other words, the standardized effect size is a way to transform the data on the results of each study into a standardized parameter, which can be analysed through statistical models. For example, the *effect* of herbivores on plant invasions can be measured in terms of the difference in total biomass of plants with and without herbivores. The larger the difference, the larger the effect size. Studies that have non-significant results will have small effect sizes, while those that have negative results will have negative effect sizes. For example, a study which found that the total biomass of plants increased with the introduction of herbivores would have a negative effect size.

Heterogeneity becomes relevant again in the next steps, when researchers assign a weight to each study, based on its quality, and estimate the *overall effect size* of all the primary studies taken together. Cases with low heterogeneity, i.e. with little variation between primary studies, are deemed 'simple'. Here, any differences in the observed effects between primary studies is assumed to be due to sampling error. Accordingly, the weight of the study is based on its sample size: the larger the sample size the higher the quality of the study and consequently, the more weight it is assigned (Dettori et al., 2022; Nakagawa et al., 2022). In statistical terms, this amounts to the inverse of the overall error variance.

In more 'complicated' cases, i.e. those with high heterogeneity between primary studies or non-independent[1] data sets, researchers use *random effects* or *multi-level* statistical models (Nakagawa & Santos, 2012). In these cases, the level of heterogeneity or non-independence affects the weighting of the primary studies: rather using the inverse of error variance, researchers use the inverse of the error variance plus the 'variance in true effects'. This, in essence, dampens the effect of the weighting, so the higher the amount of heterogeneity, the smaller the effect of the weighting. The reason for doing this is that in cases of high heterogeneity or non-independence, a higher sample size only protects against some types of

[1] Non-independence refers to a situation where the data within or between primary studies is somehow related (and thus can lead to double counting, or at least artificially magnifying the effect of some variables/relationships). In biology, this usually occurs when (i) multiple proxies are used to measure a certain trait (e.g. mating success, breeding success and survival as a proxy for fitness) or when (ii) in studies that span multiple species there is phylogenetic relatedness between a subset of these species.

bias, but not all, so our confidence in the overall effect size should not be inflated just because high sample sizes.

The amount of heterogeneity in the study pool is also reflected in the final stage of a meta-analysis, where researchers qualify the overall effect size by an 'index of precision', i.e. variance, standard error or confidence interval. In medicine, this is usually achieved through a 'Risk of Bias Assessment' which amounts to a set of statistical tests which are aimed to identify what, if any, biases can be identified in the literature as a whole and the meta-analysis in particular (Dettori et al., 2022; Konno et al., 2024). Biologists are currently developing a risk of bias framework adapted for the particularities of meta-analysis in ecology and evolution (i.e. where the levels of heterogeneity and non-independence are much higher than those in medicine, see section 4.3 for details) (Konno et al., 2024). The statistical tests offer a standardised method to interpret the results of the meta-analysis, in the sense that they can help researchers determine the confidence they should attach to the overall effect size of the meta-analysis. For example, a widely used measure of heterogeneity is I^2, which refers to the percentage of variance between effect sizes that cannot be accounted for by sampling error (Higgins & Thompson, 2002). Moreover, the widespread use of I^2 has allowed for the adoption of benchmarks, with 25%, 50%, and 75% respectively referring to small, medium, and high, heterogeneity (Senior et al., 2016).

Table 1. Summary of Evidence Synthesis

Step	Action	Type of Synthesis	Description
1	Formulate Question	Systematic Review & Meta-Analysis	- Formulate a research question, documenting and justifying decisions (e.g. on scope of the question, when and why it was revised)
2	Search Literature		- Use more than one platform (Web of Science, Scopus, Google Scholar) - Aim for effective search string, conduct backward and forward search and determine use of grey literature
3	Determine relevancy		- Define inclusion/exclusion criteria. - Depends on (sub)discipline, but types of papers that are typically discarded: - Review papers, papers on different questions, papers that do not meet certain quality thresholds (e.g. papers with insufficient/obscure data, papers with errors)
4	Calculate effect size	Meta-Analysis	- Calculate effect size from available data. - Convert to common effect size metric
5	Assign Weight		- For a fixed effect statistical model, this is based on sample size (the larger the better) - For random effects & multilevel effects statistical models, sample size matters, but its overall importance depends on amount of heterogeneity in primary research.

6	Test for bias		- Test for publication bias - Sensitivity analyses (test for explanation of heterogeneity, i.e. how much of the heterogeneity is explained by known factors?)

How much heterogeneity is typical? The answer depends on the discipline. In medicine, heterogeneity is relatively low, with 30-55% of studies having an I^2 value of 0 (that is, in 30-55% of studies there are differences in effect sizes that cannot be explained by sampling error) (Higgins & Thompson, 2002; Senior et al., 2016). This is because studies tend to focus on one species (humans), a single type of intervention (e.g. a particular drug) and similar protocols. Any heterogeneity is due to differences in the population samples (e.g. age group, geographical region, gender) or intervention procedures (e.g. different dosages). In biology, heterogeneity is typically much higher (Nakagawa & Santos, 2012; Whittaker, 2010). More specifically, Senior et al. (2016) show that ecologists should expect an I^2 of around 90%, with only 4.65% of studies having an I^2 value of 0. This is not surprising, as primary research in biology can vary in many more ways, including the very species being studied, and the method used to collect data. For example, when estimating primary productivity, the methods employed for measurement are radically different, depending on the types of vegetation being studied. Thus, in grasslands, it is usual to measure the ratio of above to below ground biomass, whereas in forests, measurements focus on above-ground biomass (the uprooting of entire trees being rather inefficient and not always ethical) (Whittaker, 2010).

2.2. The problem of heterogeneity

In very simple terms, the problem with heterogeneity is that it can make comparisons between primary studies difficult, even meaningless, effectively undermining the whole point of evidence synthesis (Stegenga, 2011). For example, if one study tests the effect of drug A on lowering blood pressure, but another tests the effect of the same drug on the rate of heart attacks, then the effect is different and so there is no way to calculate effect size. More specifically, heterogeneity between primary studies can create artificial differences between effect sizes, thus obscuring the true effect of the intervention. For example, if studies of the same drug differ in terms of dose or in terms of the time the dose is administered, then the overall effect size could be lower, thus suggesting that the drug is less effective than it actually is. Perhaps more worryingly, if only the larger dose was actually effective, but also created side effects in the patients, then pooling the studies could conceal the problem, by obscuring the percentage of patients experiencing adverse effects.

Similar arguments can be made for heterogeneity in biology, where heterogeneity is much larger (typically above 90%) (Senior et al., 2016). One of the most vociferous critiques of heterogeneity in biology is Whittaker's (2010) argument against meta-analyses of the Species-Richness-Productivity Relationship (SRPR), which, he believes, amount to 'mega-mistakes'. A meta-analysis of SRPR typically aims to determine whether higher levels of species richness contribute to higher levels of productivity. As stated in section 2.1, primary productivity can be measured in two different ways (total biomass, vs ratio of above to below ground biomass). The reason for this difference is both legitimate and unlikely to change, as the former does not require the uprooting of the entire individual – something with cannot realistically be performed on trees, only working in the context of grasses. Nonetheless, when a meta-analysis finds a difference between the productivity of grasslands and forests, is that a real difference between the two biomes or is it merely an artefact of the different methods used to measure productivity?

Whittaker argues that we cannot be sure and concludes that meta-analyses in ecology are therefore meaningless. In contrast, if we keep variation in primary studies to a minimum, then any variation in the results of primary studies will be due to real causal factors (i.e. differences in the relationship between species richness and productivity). Thus, for example, a meta-analysis where all these factors are kept constant could reveal that high levels of species richness matter more for productivity in forests than it does for grasslands. Whittaker concedes that these constraints are quite high, yet he believes that they are essential for a good meta-analysis. Moreover, he uses the stringency of the constraints as an argument against the use of meta-analysis, as he believes we simply do not have the right kind of data to conduct meta-analyses of sufficiently high quality.

Admittedly, Whittaker's paper is, by now, fifteen years old, and the rhetoric feels somewhat dated. Evidence synthesis in biology has come a long way since 2010, in the sense that it is more widespread but also more scrutinized (Gurevitch et al., 2018; Koricheva & Gurevitch, 2014; Nakagawa et al., 2017; Nakagawa & Santos, 2012). Biologists currently have more sophisticated statistical tools at their disposal (Nakagawa et al., 2022; Nakagawa & Santos, 2012), and more collective experience in conducting meta-analyses and overcoming various problems that arise (Nakagawa & Cuthill, 2007; Sánchez-Tójar et al., 2018; Sánchez-Tójar et al., 2020). Still, the high levels of heterogeneity worry even the staunch advocates of meta-analysis in biology.

There are two additional worries expressed in the literature. The first is that heterogeneity is not adequately reported in biological meta-analyses (Nakagawa & Santos, 2012; O'Connor

et al., 2017; Schielzeth & Nakagawa, 2022; Spake et al., 2022). This is problematic because it gives a false sense of security to meta-analytic results. Consider the following example (adapted from Spake et al., 2022): if a meta-analysis investigating the effect of land use change on biodiversity found an overall effect size of zero, this could be interpreted as evidence that land use change does not have an effect on biodiversity. However, this interpretation would only be correct if there was low heterogeneity between the studies, i.e. that all studies showed no (or at least non-significant) effects. If, on the other hand, there was high heterogeneity between studies, this would mean that some primary studies showed significant effects while others showed low or negative effects. In this case, we could not assume that the overall effect was representative of all cases. At the very least, we would need to conduct further investigations to determine what accounts for the heterogeneity and whether or not it could be reduced.

The second worry is that meta-analytic results with high heterogeneity might not support generalisations (Nakagawa & Cuthill, 2007; Spake et al., 2022). For example, Nakagawa & Cuthill (2007), despite advocating for the adoption of 'meta-analytic thinking' in biology, claim that "care should be taken with meta-analytic reviews in biology. Biological research can deal with a variety of species in different contexts, whereas in social and medical sciences research is centred around humans and a narrow range of model organisms, often in controlled settings. While meta-analysis of a set of similar experiments on a single species has a clear interpretation, generalization from meta-analysis across species and contexts may be questionable." (pp. 594-5). The worry seems to be that when all the primary studies in a meta-analysis focus on the same type of experiment or species, claims about that experiment or species are straightforward and legitimate. In contrast, when the meta-analysis includes data from multiple species, experimental setups etc., the overall effect size might not be equally representative of/applicable to each and every species or experimental setup. Thus, for example, a meta-analysis on the effects of fire on biodiversity that included primary research on different species, might be more representative of some communities than others, so that the overall effect (e.g. fire has no effect on biodiversity) is true of some communities (where key species have adapted to fire regimes) but not others, where there is no adaptation to fire.

3. Different goals of Evidence Synthesis

Evidence synthesis is often described as a quantitative method for amalgamating and synthesizing results from individual studies, so as to provide accurate and useful answers to the original research questions (Gurevitch et al. 2018, Koricheva et al. 2014, Stegenga 2011).

Yet if we look a bit deeper it becomes clear that syntheses can be used for different purposes. In this section I will distinguish between three such goals and explain their main differences.

3.1. Generating Causal Confidence

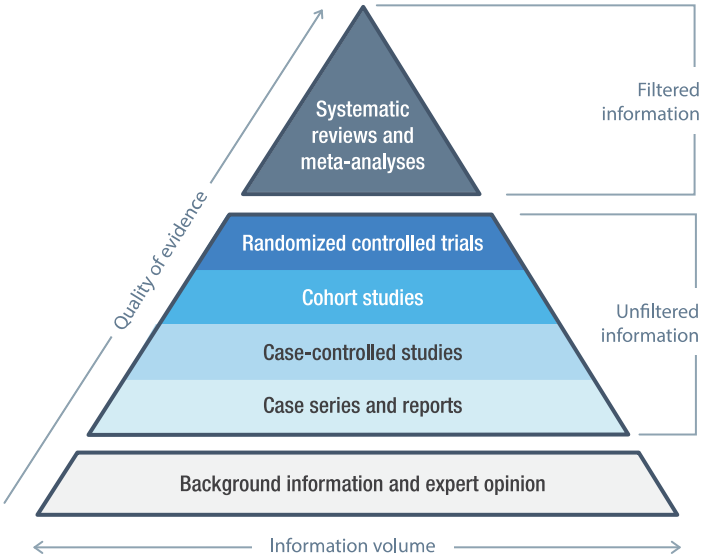
The most well-known goal of evidence synthesis is to generate or increase causal confidence. This goal pertains primarily to meta-analysis in the biomedical sciences, especially in the context of evidence-based medicine. The underlying motivation for these meta-analyses is that most primary research in medicine, e.g. randomised control trials, are necessarily limited in terms of sample size. This can be for a number of reasons, including testing of rare conditions, difficulty acquiring subjects, availability of drugs, cost of conducting the trial and so on. The problem is that with such small sample sizes, it is difficult or even impossible to definitively conclude whether or not an intervention has an effect (Egger et al., 2002; Stegenga, 2011). This is where meta-analysis comes in. It is often the case that a particular intervention is or has been tested multiple times, at different laboratories around the world. If we assume that these studies are replicates of each other, we can pool their results and generate a greater sample size, so any effect will be more likely to be statistically significant (Berlin & Golub, 2014; Carpenter, 2020; Egger et al., 2002).

For example, consider a meta-analysis that includes a number of studies on the effects of a drug on depression. A large effect size of drug *A* is meant to show that it is an effective way to tread depression. If each individual study shows a small (often not statistically significant) result, amalgamating data could provide more robust evidence of the effectiveness of the treatment. In other words, each individual study alone provides evidence of a correlation between the intervention and the effect, but together the studies are taken to indicate a true causal relationship. In addition, a meta-analysis could be used to compare the relative effectiveness of different drugs. For example, if a meta-analysis of drug *A* yields a larger effect size than a meta-analysis of drug *B*, then drug *A* is more effective for the treatment of depression.

In other words, this aim of meta-analyses is to improve on the quality of primary research, by increasing our confidence that the results of clinical trials have indeed established causal links between certain interventions and certain outcomes. A popular depiction of the quality of various types of research in evidence-based medicine is the so-called ‘pyramid of evidence’ (figure 1). At the very bottom are background information and expert opinion which are meant to provide merely anecdotal evidence. As we go up the pyramid, the quality of the studies and their results is meant to increase, while the volume of information decreases. The next three

levels are observational studies, where there is no control of confounding factors. These include case series, which are detailed observations and descriptions of individual patients, case-controlled studies, i.e. retrospective studies where researchers compare existing observations of a number of patients, and cohort studies, where patients are followed and observed over a period of time. At the top of the primary research section (which in medicine is often referred to as *unfiltered information*) are randomised control trials (RCTs), where patients are randomized and placed in the treatment or control groups, the latter of which receive a placebo rather than the treatment. RCTs aim to control for confounding variables and identify genuine causal links between the intervention and the outcome. Given the position of RCTs in the pyramid they often referred to as the 'gold standard' of evidence (Stegenga, 2011). For advocates of the pyramid, systematic reviews and meta-analyses are the pinnacle of the pyramid as they filter the information provided by RCTs, along with eliminating or at least minimizing their deficits, thus constituting the 'platinum standard of evidence' (see discussion in Stegenga, 2011).

Figure 1. Pyramid of evidence



A typical evidence pyramid in evidence-based medicine. As we go up the pyramid, the volume of information decreases, and the quality of evidence increases. Unfiltered information refers to primary studies and filtered information refers to meta-level research, i.e. systematic reviews and meta-analyses. From (Tannenbaum & Sebastian, 2022).

3.2. Arbitrating between Contradictory Results

It is often the case that primary research on a certain topic differs. In clinical trials, for example, one study might find that a certain intervention has a significant effect, while another trial might find a non-significant effect. Sometimes results can be downright contradictory,

with some studies finding a positive relationship between two variables and others finding a negative relationship between the same variables. Meta-analyses can be used to help researchers determine how to deal with varying or contradictory results by providing an overall assessment of the effect. One way to achieve this is through the process of weighting (see section 2.1). For example, a meta-analysis could reveal that the studies showing no effects of drug *A*, have small sample sizes and should be weighted less heavily. In other words, by providing a judgement on the relative quality of various studies, a meta-analysis can show that some apparent contradictions between primary studies can easily be resolved.

This process of weighting the primary studies is considered to be a significant advance for evidence synthesis, and one of the main reasons to prefer meta-analysis to its predecessors, especially 'vote counting' (Koricheva et al., 2013; Nakagawa & Poulin, 2012). This involves sorting primary research into three categories (significant results in favour of hypothesis, significant results against hypothesis and non-significant results), determining which category has the highest number of studies, and declaring that category the 'winner' (Koricheva & Gurevitch, 2013). A major problem with vote counting that it cannot take into account the quality of the primary studies, giving equal weight to high and low-quality studies i.e., those with low sample sizes. This leads to biased and misleading results at the meta-research level, which has been extensively documented (Koricheva & Gurevitch, 2013; Nakagawa et al., 2017; Nakagawa & Poulin, 2012).

Meta-analyses can also reveal how different measurements of a certain phenomenon can lead to different conclusions, while also providing information about how to deal with the resulting contradictory conclusions. Consider the case of biodiversity trends, i.e. whether biodiversity is increasing or decreasing in the last decades. While many studies have concluded that biodiversity is decreasing, there have been some studies which demonstrate an increase in biodiversity. This is interesting but also potentially problematic as it can have an effect on conservation policy and funding allocation, as it can be used as 'evidence' for decreasing the funding allocated to conservation efforts (Fieseler, 2021; Pyron, 2017). In a meta-analysis of biodiversity trends in Europe, Pilotto et al., (2020) found that many of the studies which found no changes or increases in biodiversity were measuring species turnover rather than species richness or abundance. These are instances where the overall number of species might be increasing, but this is due to biological invasions, i.e. the native species are actually being replaced by alien species. Thus, the meta-analysis showed that if we are interested in conservation of native species in Europe, we can discount the studies that measure species turnover.

3.3. Exploring the Scope of Generalisations

Perhaps the least well-known, but in my opinion, the most useful goal of meta-analysis is a tool for testing the scope of generalisations. At first glance, it seems similar to the goal outlined in the previous section, as it also is a way to deal with differing or contradictory primary results. However, there is a subtle but important difference between the two goals. Here, a meta-analysis is not used to determine which side of the primary research 'wins', rather it is used to determine when or where a causal connection between two variables holds and when or where it breaks down.

Consider, for example, the case of the 'enemy release hypothesis' in invasion biology. The basic idea is quite simple: alien species do not encounter their traditional enemies in new territories, so they can thrive. The situation becomes trickier when scientists tried to determine how exactly enemy release manifests in each case, and what conclusions can be drawn from it (Heger & Jeschke, 2014). For example, while there have been documented cases where alien plants or their seeds are not consumed by native predators, there are also number of cases where alien plants actually attracted native herbivores, and that these herbivores have a significant negative effect on seed production and plant survival (both of which seem to be quite important for a successful invasion) (Maron & Vilà, 2001). Studies at different scales also tend to yield contradictory results, as larger-scale biogeographical analyses primarily show a reduction in the diversity of enemies in the introduced range compared with the native range, while smaller-scale community studies often show that alien species are no less affected by enemies than native species in the invaded community (Colautti et al., 2004).

A meta-analysis conducted in 2006 revealed some interesting insights regarding these contradictory results. Parker et al., (2006) analysed 63 manipulative field studies of plant invasions which included the effect of herbivores on the outcome of the invasion (i.e. they included primary studies where herbivores facilitated and where they hindered the plant invasion). At first glance, it seemed that there was stronger evidence against the enemy release hypothesis, as there were cases where native herbivores decreased the abundance of alien plants, i.e. plants encountered new enemies, and cases where alien herbivores i.e. their existing enemies increased the abundance of alien plants. However, they also found that the negative effect of native herbivores on the alien plants was weaker than the positive effect of alien herbivores on them (28% reduction in the former vs 65% increase in the latter). Probing deeper, they realised that some studies focused on invertebrate herbivores while others focused on

vertebrates. It turns out that native vertebrate herbivores had a three to five-fold larger negative impact on alien plant survival than native invertebrate herbivores.

What accounts for this difference in the strength of the effect across studies? A closer look at the primary research revealed that the native invertebrate herbivores were specialists (i.e. they prey on specific plant species) while the alien vertebrate herbivores were generalists (i.e. they prey indiscriminately on many different plant species). This is the final piece of the puzzle, which explains the apparent contradictions by showing the limits of the enemy release hypothesis. In other words, the enemy release mechanisms function normally in cases the native herbivores are specialists and there are no alien herbivores; here the alien plants are released from their old enemies but are not affected by the native specialists, who continue to focus on their preferred native plants. However, the enemy release effect is counteracted (or at least overshadowed) by the existence of alien generalist predators, who consume both native and alien plants. In fact, these generalist alien predators might, in some cases, prefer the native plants, thus further facilitating the spread of the alien plant invaders.

I believe that this is an extremely useful way to utilize evidence synthesis. One of the main problems in ecology is the difficulty of constructing generalisations that can support explanations and predictions (Beckage et al., 2011; Houlahan et al., 2017; Kaunisto et al., 2016; Lawton, 1999; Mitchell, 2002; Raerinne, 2014; Turchin, 2001). More specifically, while ecologists are able to identify patterns in the phenomena they study, these patterns often break down (Elliott-Graves 2024; Doak et al., 2008). This means that ecological generalisations are often limited in scope (Elliott-Graves 2024; Mitchell, 2000). This creates problems for ecological research, as generalisations form the basis for some types of explanations and most predictions; a generalisation breaking down translates into knowledge not being transferrable across systems or across time periods (Catford et al., 2022; Spake et al., 2023). While ecologists generally aware of these issues, they are nonetheless extremely challenging, especially in applied contexts, when ecologists only have a little time and few options to intervene on a system (Catford et al., 2022; Doak et al., 2008; Mouquet et al., 2015). Thus, any information on the scope and limits of a generalisation can be incredibly useful; it can make the difference between a successful and unsuccessful intervention. In the case of enemy release, for example, knowing that the enemy release effect is overshadowed by generalist herbivores can have important effects on policy. Thus, scientists aiming to save a native plant from extinction should not merely focus on predation from local insects, rather they should focus predominantly on shielding the plant from non-native herbivores.

The discussion in this section was intended to foreshadow the idea that the effects of heterogeneity are not uniform but can differ depending on the goal of the synthesis in question. In the next section, I will examine the cases where heterogeneity is problematic.

4. When is Heterogeneity a genuine problem?

Most cases where heterogeneity is genuinely problematic occur when our expectations of heterogeneity do not match reality, that is, when there is (much) more heterogeneity than we expected. This can occur when heterogeneity is unreported, which happens when a synthesis contains no (or insufficient) information about the heterogeneity of primary studies included in the synthesis. The problem is that unreported heterogeneity implies that there is no significant heterogeneity in the studies, so any overall effect sizes are taken at face value. If, however, there is significant heterogeneity in the effect size, then the issues outlined in section 2, hold, that is, we cannot be sure that the overall effect size accurately represents the pool of primary studies (Nakagawa et al., 2017; Spake et al., 2022). Moreover, lack of information about heterogeneity can also hamper subsequent efforts to correct or further investigate the possible effects of heterogeneity as novel statistical tools are developed (Ioannidis et al., 2007; Senior et al., 2016).

A particularly pernicious set of cases where heterogeneity does not match expectations, occurs when the synthesis in question is used for the goal outlined in section 3.1., namely 'generating causal confidence'. Recall that this use of meta-analysis involves pooling results from different studies in the hope of generating a larger overall effect size, thus demonstrating a stronger link between the intervention and the effect. Here, researchers treat the primary studies as though they were replicates of each other, i.e., they assume that there is a high level of homogeneity between the studies, so that any differences between control and experimental groups can be safely attributed to the intervention itself. However, if it turns out that heterogeneity between primary studies is high, then we cannot be sure that the variation is attributable to the intervention and the very premise of the meta-analysis is undermined.

The issue is that heterogeneity between primary studies can create artificial differences between effect sizes. Consider again the example outlined in section 2, where the meta-analysis is aiming to show a significant effect size for a certain drug. However, primary studies differ in terms of the dosage administered. In this example, only the higher dose of the drug is actually effective yet also creates side effects in a subset of the patients. Assuming *homo*geneity and pooling the results obscures both these important issues. First, it fails to show that different dosages have different effects but implies that a median dosage has a sufficient effect. Second,

it obscures the connection between the higher dosage and the side effects. In other words, assuming homogeneity and pooling the results, dilutes the variation between primary studies and obscures issues that ought to be highlighted.

5. Can Heterogeneity be valuable?

While it is undeniable that heterogeneity is problematic in the contexts described in the previous section, I believe that there are other contexts where it is much less detrimental, and even cases where it can be useful. What I mean when I say that heterogeneity is often less detrimental than we expect, is that there are numerous statistical tools for accounting for and addressing heterogeneity. A number of scholars explicitly state that heterogeneity is not problematic per se, but only becomes problematic if it is unexpected, un(der)reported or un(der)investigated (Higgins, 2008; Higgins & Thompson, 2002; Nakagawa et al., 2017; Schielzeth & Nakagawa, 2022; Senior et al., 2016 see also discussion in section 4). When heterogeneity is expected and adequately reported, then researchers have access to numerous methods for further investigating the causes of heterogeneity along with its effects (Senior et al., 2016). For instance, it is becoming standard practice in biological meta-analyses to use random effects models or mixed effects models, which help researchers analyse heterogeneity rather than fixed effects models, which assume low levels of heterogeneity (Senior et al., 2016). Mixed effects models allow heterogeneity to be partitioned, so that it is possible to distinguish between possible causes of heterogeneity, such as phylogenetic heritability in multi-species studies (Senior et al., 2016). Of course, many of these tests are time-consuming and require some statistical knowledge, yet they are usually readily available and free.[2]

But how exactly can analysing heterogeneity be useful? Unlike the context of generating causal confidence, when we are using meta-analyses to arbitrate between contradictory results (3.2), or examine the scope of generalisations (3.3), heterogeneity can provide us with valuable information. Starting with the case of contradictory results, heterogeneity is useful when groups of primary studies emerge which display intra-group homogeneity and inter-group heterogeneity, in other words, when heterogeneity clusters in interesting ways. For example, in the meta-analysis of biodiversity trends in Europe, Pilotto et al. (2020) found that the primary studies clustered in terms of how biodiversity was measured: the studies which showed decreases in biodiversity were those that measured richness or abundance whereas those that

[2] Most of these tests can be easily implemented by running existing software packages in *R.* In my experience, many of the biologists who have developed/adapted these packages for biological data are also extremely helpful, willing to answer questions and troubleshoot the implementation of the software.

showed no changes or increases in biodiversity were those that measured species turnover. These heterogeneous clusters are thus quite informative when we are trying to make sense of contradictory results. In this case, they show us that biodiversity of native species is decreasing in Europe and that any increases in biodiversity are due to invasive species. This means that, rather than being reassured from any results that show increases in biodiversity, we should expand our conservation strategies to include management of invasive species. In other words, the clustering shows us that any contradiction between results is, at least from a conservation standpoint, illusory.

Clusters of heterogeneity can also be informative in the sense that they can uncover biases in certain experimental setups, measurements or species. In the case of medicine, for example, if results cluster by geographical region or dosage then this is an indication that there is something about how the experiment was conducted in certain contexts which could account for the different results. In the case of biology, if the clusters correlate to particular species, for example, this could indicate that there is something problematic with the measurement of the effect in that species. Of course, it could indicate that there is a real difference in effect in that species – I will discuss this point in the next paragraph. The point is that heterogeneous clusters, if adequately investigated, can account for contradictory results, and can even provide additional information which explains the underlying causes of the contradiction.

The case for preserving and analysing heterogeneity is even stronger in the context of exploring the scope of generalisations, as it is the existence of heterogeneity itself that predicts the limits of a generalisation and in some cases can even help in explaining the limits of the generalisation in question. I will return to the case of the enemy release hypothesis, outlined in section 3.3. Here, the scientists were able to explain the reason why primary studies examining the enemy release hypothesis yielded contradictory results, as they realised that the enemy release mechanism is sometimes overshadowed by other mechanisms (those generated by generalist herbivores). Thus, the heterogeneity in the primary studies provided important information about scope of the enemy release hypothesis, i.e. where then mechanism of enemy release was effective and where it was not. In fact, in this case, if the researchers were to reduce the heterogeneity of their sample in the traditionally approved way, i.e. by excluding the studies on one type of herbivore (i.e. insects or vertebrates), they would have missed two important insights.

First, they would not have realised that the key difference regarding enemy release was in terms of whether the herbivores were specialists or generalists (which happened to coincide with the categories of insect and vertebrate). If they had excluded one group by default, they

would not have realised the limit in scope of the enemy release mechanism, i.e. when it was overshadowed by other mechanisms. Second, failing to reach this conclusion would also have prevented the scientists from another insight into biological invasions, namely that this explains another perplexing phenomenon, namely why it is much more common for European plants to invade areas outside Europe, rather than vice versa. The answer is that generalist herbivores from Europe, such as pigs, horses and cattle, are more widespread than generalist herbivores from other continents and contribute more often to the success of exotic plants with which they have co-evolved.

In short, 'correcting' for heterogeneity, i.e. leaving out the primary studies that increase the heterogeneity of the overall effect size can create more problems than it solves. Here, heterogeneity is a feature rather than bug, and though all heterogeneity should be investigated, it should not automatically be met with suspicion. Indeed, some researchers argue that in disciplines with expectations of high heterogeneity, such as biology, it is instances of low heterogeneity that should be treated with suspicion or at least subjected to similar amounts of scrutiny as cases of high heterogeneity (Senior et al., 2016).

Most of this discussion pertains to disciplines, such as biology, where high heterogeneity is expected. Moreover, in section 4, I argued that heterogeneity is indeed problematic when it is higher than expected, which is usually the case in medicine. But are there contexts in which high heterogeneity can also be useful in medicine? I believe that with some conceptual and methodological modifications to medical meta-analyses, heterogeneity could also be informative here. Thus, if medical meta-analysts adopted goals 2 or 3 rather than 1, along with the appropriate statistical methods for analysing heterogeneity, then it could potentially be as useful as it is in biology. More specifically, if a meta-analysis was intended to explore the limits of a causal claim, rather than aiming to generate causal confidence, then the statistical tools could be used to reveal clusters that could potentially be informative. For example, if the results cluster in terms of age group, sex, additional health issues etc., this could show that the particular drug only works on say, men between the ages of 18-60, that a particular dosage has serious side effects on post-menopausal women and so on.[3] The important

[3] I should note that here, as we are dealing with a single species, it is still possible to 'pool' results at the same time as investigating heterogeneity clusters. That is, if meta-analysists have information on sex, age, additional health issues etc., from the primary studies (something which is increasingly the case) then they can pool the subgroups from different studies to test whether the population as a whole, clusters in interesting ways. The beauty of these statistical tools is that once all the data is put into the program, rearranging it in different clusters is a matter of seconds.

point, yet again, is not whether heterogeneity exists or not, but how it is approached and how it is analysed.

6. Conclusion

Even though exploring the limits of a hypothesis is primarily associated with evidence synthesis biology, it can also be used in other disciplines, including medicine. Thus, for example, researchers could (re)analyse primary data on different sexes, age groups etc., and identify limitations in scope for medications, or even identify consistent bias in diagnoses. The point is that what determines whether heterogeneity is a problem depends on our attitude towards heterogeneity. If we assume that it is non-existent, when it does exist, then our synthesis will suffer. If, on the other hand, we expect some heterogeneity to exist and explicitly analyse it, then we can end up with a lot more information than we would from an entirely homogeneous set of primary research (Higgins & Thompson, 2002; Spake et al., 2022). To sum up, heterogeneity is here to stay, but this does not seem to be the insurmountable problem that early critics claimed it was. The availability of new and easily implementable statistical packages, make exploring heterogeneity and integral but also useful dimension of evidence synthesis.

7. Bibliography

Beckage, B., Gross, L. J., & Kauffman, S. (2011). The limits to prediction in ecological systems. *Dx.Doi.Org*, *2*(11), art125. https://doi.org/10.1890/ES11-00211.1

Berchialla, P., Chiffi, D., Valente, G., & Voutilainen, A. (2020). The power of meta-analysis: A challenge for evidence-based medicine. *European Journal for Philosophy of Science*, *11*(1), 7. https://doi.org/10.1007/s13194-020-00321-w

Berlin, J. A., & Golub, R. M. (2014). Meta-analysis as Evidence: Building a Better Pyramid. *JAMA*, *312*(6), 603–606. https://doi.org/10.1001/jama.2014.8167

Booth, A., Clarke, M., Dooley, G., Ghersi, D., Moher, D., Petticrew, M., & Stewart, L. (2012). The nuts and bolts of PROSPERO: An international prospective register of systematic reviews. *Systematic Reviews*, *1*(1), 2. https://doi.org/10.1186/2046-4053-1-2

Bruner, J. P., & Holman, B. (2019). Self-correction in science: Meta-analysis, bias and social structure. *Studies in History and Philosophy of Science Part A*, *78*, 93–97. https://doi.org/10.1016/j.shpsa.2019.02.001

Cadotte, M. W., Mehrkens, L. R., & Menge, D. N. L. (2012). Gauging the impact of meta-analysis on ecology. *Evolutionary Ecology*, *26*(5), 1153–1167. https://doi.org/10.1007/s10682-012-9585-z

Carpenter, C. J. (2020). Meta-Analyzing Apples and Oranges: How to Make Applesauce Instead of Fruit Salad. *Human Communication Research*, *46*(2–3), 322–333. https://doi.org/10.1093/hcr/hqz018

Catford, J. A., Wilson, J. R. U., Pyšek, P., Hulme, P. E., & Duncan, R. P. (2022). Addressing context dependence in ecology. *Trends in Ecology & Evolution*, *37*(2), 158–170. https://doi.org/10.1016/j.tree.2021.09.007

Chen, H., & Jhanji, V. (2012). Survey of systematic reviews and meta-analyses published in ophthalmology. *The British Journal of Ophthalmology*, *96*, 896–899. https://doi.org/10.1136/bjophthalmol-2012-301589

Colautti, R. I., Ricciardi, A., Grigorovich, I., & MacIsaac, H. J. (2004). Is invasion success explained by the enemy release hypothesis? *OIKOS*, *7*, 721–733.

Dettori, J. R., Norvell, D. C., & Chapman, J. R. (2022). Fixed-Effect vs Random-Effects Models for Meta-Analysis: 3 Points to Consider. *Global Spine Journal*, *12*(7), 1624–1626. https://doi.org/10.1177/21925682221110527

Doak, D. F., Estes, J. A., Halpern, B. S., Jacob, U., Lindberg, D. R., Lovvorn, J., Monson, D. H., Tinker, M. T., Williams, T. M., Wootton, J. T., Carroll, I., Emmerson, M., Micheli, F., & Novak, M. (2008). Understanding and predicting ecological dynamics: Are major surprises inevitable? *Ecology*, *89*(4), 952–961. https://doi.org/10.1890/07-0965.1

Egger, M., Ebrahim, S., & Smith, G. D. (2002). Where now for meta-analysis? *International Journal of Epidemiology*, *31*(1), 1–5. https://doi.org/10.1093/ije/31.1.1

Elliott-Graves, A. (2023) Ecological Complexity. Cambridge University Press. https://doi.org/10.1017/9781108900010

Fieseler, C. (2021). The case against the concept of biodiversity. *Vox*. https://www.vox.com/22584103/biodiversity-species-conservation-debate

Fletcher, S. C. (2022). Replication Is for Meta-Analysis. *Philosophy of Science*, *89*(5), 960–969. https://doi.org/10.1017/psa.2022.38

Fontelo, P., & Liu, F. (2018). A review of recent publication trends from top publishing countries. *Systematic Reviews*, *7*(1), 147. https://doi.org/10.1186/s13643-018-0819-1

Foo, Y. Z., O'Dea, R. E., Koricheva, J., Nakagawa, S., & Lagisz, M. (2021). A practical guide to question formation, systematic searching and study screening for literature reviews in ecology and evolution. *Methods in Ecology and Evolution*, *12*(9), 1705–1720. https://doi.org/10.1111/2041-210X.13654

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, *555*(7695), 175–182. https://doi.org/10.1038/nature25753

Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. P. A. (2020). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*, *7*(1), 11–37. https://doi.org/10.1146/annurev-statistics-031219-041104

Heger, T., & Jeschke, J. (2014). The enemy release hypothesis as a hierarchy of hypotheses. *Oikos*, *123*(6), 741–750. https://doi.org/10.1111/j.1600-0706.2013.01263.x

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, *37*(5), 1158–1160. https://doi.org/10.1093/ije/dyn204

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Holman, B. (2019). In defense of meta-analysis. *Synthese*, *196*(8), 3189–3211. https://doi.org/10.1007/s11229-018-1690-2

Houlahan, J., McKinney, S., Anderson, M., & McGill, B. (2017). The priority of prediction in ecological understanding. *Oikos*, *126*(1), 1–7. https://doi.org/10.1111/oik.03726

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2016). The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *The Milbank Quarterly*, *94*(3), 485–514. https://doi.org/10.1111/1468-0009.12210

Ioannidis, J. P. A., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*, *335*(7626), 914–916. https://doi.org/10.1136/bmj.39343.408449.80

Jukola, S. (2017). On ideals of objectivity, judgments, and bias in medical research—A comment on Stegenga. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *62*, 35–41. https://doi.org/10.1016/j.shpsc.2017.02.001

Kaunisto, S., Ferguson, L. V., & Sinclair, B. J. (2016). Can we predict the effects of multiple stressors on insects in a changing climate? *Current Opinion in Insect Science*, *17*, 55–61. https://doi.org/10.1016/j.cois.2016.07.001

Konno, K., Gibbons, J., Lewis, R., & Pullin, A. S. (2024). Potential types of bias when estimating causal effects in environmental research and how to interpret them. *Environmental Evidence*, *13*(1), 1. https://doi.org/10.1186/s13750-024-00324-7

Koricheva, J., & Gurevitch, J. (2013). Place of Meta-analysis among other Methods of research synthesis. In *Handbook of Meta-Analysis in Ecology and Evolution*. Princeton University Press.

Koricheva, J., & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology*, *102*(4), 828–844. https://doi.org/10.1111/1365-2745.12224

Koricheva, J., Gurevitch, J., & Mengersen, K. (2013). *Handbook of Meta-analysis in Ecology and Evolution*. Princeton University Press.

Kovaka, K. (2022). Meta-Analysis and Conservation Science. *Philosophy of Science*, *89*(5), 980–990. https://doi.org/10.1017/psa.2022.68

Lawton, J. H. (1999). Are there general laws in ecology? *Oikos*, *84*(2), 177–192. https://doi.org/10.2307/3546712

Maron, J. L., & Vilà, M. (2001). When do herbivores affect plant invasion? Evidence for the natural enemies and biotic resistance hypotheses. *Oikos*, *95*(3), 361–373. https://doi.org/10.1034/j.1600-0706.2001.950301.x

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in Second Language Research: Narrative and Systematic Reviews and Recommendations for the Field. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286

Maziarz, M. (2022). Is meta-analysis of RCTs assessing the efficacy of interventions a reliable source of evidence for therapeutic decisions? *Studies in History and Philosophy of Science*, *91*, 159–167. https://doi.org/10.1016/j.shpsa.2021.11.007

Mitchell, S. D. (2000). Dimensions of Scientific Law. *Philosophy of Science*, *67*(2), 242–265. https://doi.org/10.2307/188723?refreqid=search-gateway:148949eeb4f88afdc4c4ac3f69c73275

Mitchell, S. D. (2002). Ceteris Paribus—An Inadequate Representation For Biological Contingency. *Erkenntnis*, *57*(3), 329–350.

Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputié, A., Eveillard, D., Faure, D., Garnier, E., Gimenez, O., Huneman, P., Jabot, F., Jarne, P., Joly, D., Julliard, R., Kéfi, S., Kergoat, G. J., Lavorel, S., Le Gall, L., Meslin, L., … Loreau, M. (2015). Predictive ecology in a changing world. *Journal of Applied Ecology*, *52*(5), 1293–1310. https://doi.org/10.1111/1365-2664.12482

Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2019). Replication studies in economics—How many and which papers are chosen for replication, and why? *Research Policy*, *48*(1), 62–83.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*(4), 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x

Nakagawa, S., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W. A., Parker, T. H., Sánchez-Tójar, A., Yang, Y., & O'Dea, R. E. (2022). Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods in Ecology and Evolution*, *13*(1), 4–21. https://doi.org/10.1111/2041-210X.13724

Nakagawa, S., Noble, D. W. A., Senior, A. M., & Lagisz, M. (2017). Meta-evaluation of meta-analysis: Ten appraisal questions for biologists. *BMC Biology*, *15*(1), 18. https://doi.org/10.1186/s12915-017-0357-7

Nakagawa, S., & Poulin, R. (2012). Meta-analytic insights into evolutionary ecology: An introduction and synthesis. *Evolutionary Ecology*, *26*(5), 1085–1099. https://doi.org/10.1007/s10682-012-9593-z

Nakagawa, S., & Santos, E. S. A. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, *26*(5), 1253–1274. https://doi.org/10.1007/s10682-012-9555-5

O'Connor, M. I., Gonzalez, A., Byrnes, J. E. K., Cardinale, B. J., Duffy, J. E., Gamfeldt, L., Griffin, J. N., Hooper, D., Hungate, B. A., Paquette, A., Thompson, P. L., Dee, L. E., & Dolan, K. L. (2017). A general biodiversity–function relationship is mediated by trophic level. *Oikos*, *126*(1), 18–31. https://doi.org/10.1111/oik.03652

Pilotto, F., Kühn, I., Adrian, R., Alber, R., Alignier, A., Andrews, C., Bäck, J., Barbaro, L., Beaumont, D., Beenaerts, N., Benham, S., Boukal, D. S., Bretagnolle, V., Camatti, E., Canullo, R., Cardoso, P. G., Ens, B. J., Everaert, G., Evtimova, V., … Haase, P. (2020). Meta-analysis of multidecadal biodiversity trends in Europe. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17171-y

Pyron, R. A. (2017). We don't need to save endangered species. Extinction is part of evolution. *Washington Post*. https://www.washingtonpost.com/outlook/we-dont-need-to-save-endangered-species-extinction-is-part-of-evolution/2017/11/21/57fc5658-cdb4-11e7-a1a3-0d1e45a6de3d_story.html

Raerinne, J. (2014). Evolutionary Contingency, Stability, and Biological Laws. *Journal for General Philosophy of Science*, *46*(1), 45–62. https://doi.org/10.1007/s10838-014-9271-7

Romero, F. (2016). Can the behavioral sciences self-correct? A social epistemic study. *Studies in History and Philosophy of Science Part A*, *60*, 55–69. https://doi.org/10.1016/j.shpsa.2016.10.002

Sánchez-Tójar, A., Moran, N. P., O'Dea, R. E., Reinhold, K., & Nakagawa, S. (2020). Illustrating the importance of meta-analysing variances alongside means in ecology and evolution. *Journal of Evolutionary Biology*, jeb.13661. https://doi.org/10.1111/jeb.13661

Sánchez-Tójar, A., Nakagawa, S., Sánchez-Fortún, M., Martin, D. A., Ramani, S., Girndt, A., Bókony, V., Kempenaers, B., Liker, A., Westneat, D. F., Burke, T., & Schroeder, J. (2018). Meta-analysis challenges a textbook example of status signalling and demonstrates publication bias. *eLife*, *7*, e37385. https://doi.org/10.7554/eLife.37385

Schielzeth, H., & Nakagawa, S. (2022). Conditional repeatability and the variance explained by reaction norm variation in random slope models. *Methods in Ecology and Evolution*, *13*(6), 1214–1223. https://doi.org/10.1111/2041-210X.13856

Senior, A. M., Grueber, C. E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E. S. A., & Nakagawa, S. (2016). Heterogeneity in ecological and evolutionary meta-analyses: Its magnitude and implications. *Ecology*, *97*(12), 3293–3299. https://doi.org/10.1002/ecy.1591

Spake, R., Bowler, D. E., Callaghan, C. T., Blowes, S. A., Doncaster, C. P., Antão, L. H., Nakagawa, S., McElreath, R., & Chase, J. M. (2023). Understanding 'it depends' in ecology: A guide to hypothesising, visualising and interpreting statistical interactions. *Biological Reviews*, *98*(4), 983–1002. https://doi.org/10.1111/brv.12939

Spake, R., O'Dea, R. E., Nakagawa, S., Doncaster, C. P., Ryo, M., Callaghan, C. T., & Bullock, J. M. (2022). Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology & Evolution*, *6*(12), 1818–1828. https://doi.org/10.1038/s41559-022-01891-z

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Biol & Biomed Sci*, *42*(4), 497–507. https://doi.org/10.1016/j.shpsc.2011.07.003

Tannenbaum, M., & Sebastian, S. (2022). *Levels of Evidence*. OpenMD Reference Guides. https://openmd.com/guide/levels-of-evidence

Taylor, A., & Munafò, M. (2016). Triangulating meta-analyses: The example of the serotonin transporter gene, stressful life events and major depression. *BMC Psychology*, *4*. https://doi.org/10.1186/s40359-016-0129-0

Turchin, P. (2001). Does Population Ecology Have General Laws? *Oikos*, *94*(1), 17–26.

Watkins, H. V., Yan, H. F., Dunic, J. C., & Côté, I. M. (2021). Research biases create overrepresented "poster children" of marine invasion ecology. *Conservation Letters*, *14*(3), e12802. https://doi.org/10.1111/conl.12802

Whittaker, R. J. (2010). Meta-analyses and mega-mistakes: Calling time on meta-analysis of the species richness–productivity relationship. *Ecology*, *91*(9), 2522–2533. https://doi.org/10.1890/08-0968.1