Philosophy of Science (2026), **93**, 541–560
doi:10.1017/psa.2025.10176

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

ARTICLE

Reevaluating Heterogeneity in Evidence Synthesis

Alkistis Elliott-Graves

Department of Philosophy, Bielefeld University, Bielefeld, Germany
Email: a.elliott-graves@uni-bielefeld.de

(Received 07 April 2025; revised 05 September 2025; accepted 07 September 2025; first published online 23 September 2025)

Abstract

Evidence synthesis has recently seen an enormous increase across many different scientific disciplines. Despite its popularity, it has also been the subject of significant criticism. One of the main critiques of evidence synthesis is the existence and treatment of *heterogeneity* between primary studies. The aim of this article is to reexamine heterogeneity in evidence synthesis, including perspectives from evolutionary biology, ecology, and conservation. I argue that while some of the critiques of heterogeneity remain valid, there are contexts where it is much less problematic. Furthermore, I propose that heterogeneity can be useful because analyzing it can provide valuable information.

1. Introduction

In many scientific disciplines, including medicine, psychology, and biology, 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 identifying connections between causes and effects or actively manipulating systems to bring about certain effects. However, primary studies usually differ in a number of ways: whether they are observational or experimental, whether the experiment is in a laboratory or in the field, which variables are being measured, the type and size of the population being studied, and so on. So, replicating the results of a study is not always feasible (Hardwicke et al. 2020; Koricheva et al. 2013; Marsden et al. 2018; Mueller-Langer et al. 2019). Moreover, these differences between primary studies often make comparing their results difficult. Scientists are increasingly using "evidence synthesis," which primarily consists of *systematic review* and *meta-analysis*, to integrate and synthesize results from individual studies, so as to provide general answers to the original research questions, as well as information about the scope of their results (Gurevitch et al. 2018; Koricheva et al. 2013; 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 articles between the 1990s and the 2010s (see, e.g., Cadotte et al. 2012; Chen and Jhanji 2012; Fontelo and Liu 2018; Taylor and Munafò 2016). Perhaps unsurprisingly, the

© The Author(s), 2025. Published by Cambridge University Press on behalf of Philosophy of Science Association.

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

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 their peers not to rely on these methods because they would not only fail at their intended goals (e.g., of minimizing bias and helping to overcome the replication crisis) but could also actively lower the overall quality of research in a field. The two main critiques are (i) that evidence synthesis perpetuates existing biases and introduces new types of bias (Ioannidis 2005, 2016; Romero 2016; Stegenga 2011; Watkins et al. 2021) and (ii) 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).

Philosophers have begun to enter the furor surrounding evidence synthesis (Berchialla et al. 2020; Bruner and Holman 2019; Fletcher 2022; Holman 2019; Jukola 2017; Kovaka 2022; LaCaze and Osimani 2020; Maziarz 2022; Stegenga 2011; Worrall 2002). Yet the philosophical coverage of the various topics is not comprehensive because most articles focus on evidence synthesis in the context of medicine and parts of psychology (Berchialla et al. 2020; Bruner and Holman 2019; Fletcher 2022; Holman 2019; Jukola 2017; LaCaze and Osimani 2020; Maziarz 2022; Worrall 2002). In addition, many philosophical accounts 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 that take into account evidence synthesis in *biology* as well as medicine and the social sciences.

The aim of this article is to reexamine the issue of heterogeneity in evidence synthesis. While the original critiques of heterogeneity highlight some valid points, I will argue that these points are mostly 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 generalizations*, 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 it can sometimes 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 some key terms and processes in evidence synthesis (sec. 2). I then delve deeper into the notion of heterogeneity and outline the main critiques leveled against it (sec. 3). I argue that evidence synthesis is used for different purposes, which usually align with different disciplines (e.g., medicine vs. biology; sec. 4). In section 5, I explain when and why heterogeneity is genuinely problematic, and in section 6, I contrast these cases with some where heterogeneity can be valuable. Section 7 provides some concluding remarks.

2. Evidence synthesis: Key terms and procedures

The terminology surrounding evidence synthesis can be confusing because different bodies of literature have emerged within each discipline that uses it. While the

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

Cochrane Handbook is often cited as the final word in terms of definitions, there are some terms that are discipline specific (Gurevitch et al. 2018; Nakagawa et al. 2020; Siddaway et al. 2019). In the philosophical literature, most articles adopt the terminology from various branches of medicine or psychology, yet some philosophers have coined new terms to describe aspects of "meta-research" (see, e.g., Osimani 2020 and table 1). Because this article focuses primarily on evidence synthesis in biology, I will be following the terminology of that field. I should note that this terminology (i.e., in the biology literature) is quite comprehensively thought-out; there are a number of articles that explain and justify key terms, and the majority of these terms follow, or at least are compatible with, the *Cochrane Handbook*'s terminology (Gurevitch et al. 2018; Nakagawa et al. 2020). Finally, this terminological usage is widespread and relatively consistent within the biological evidence synthesis literature.

To avoid further confusion, I have summarized some key terms in table 1, indicating relevant references, along with some clarifications that may be helpful to the reader.

Following the biology literature, I will use *evidence synthesis* as the term that includes systematic review and meta-analysis. These two procedures 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).

3. Heterogeneity and its effects

The main source of heterogeneity in evidence synthesis is the primary research on which the synthesis is based. Primary studies differ in many ways, in terms of their inputs (e.g., experimental setup, type of intervention, length of treatment, phenomenon/species/taxon being studied) and outputs (e.g., effect size, magnitude/direction of effect, how the effect is measured/presented). The more variation there is in the sample of articles analyzed, the more heterogeneity there is in the synthesis. In addition, the larger the basic pool of articles is, the more heterogeneous they are likely to be because a larger pool of articles 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 in outputs (effect size, magnitude/direction of effect, etc.). Still, limiting the number of studies too much is also dangerous because it can increase the risk of bias (e.g., publication bias, which occurs because positive results tend to be published more often than neutral or negative results; Sánchez-Tójar et al. 2018). Thus, researchers aim to find the "sweet spot" so that their synthesis is broad enough to reduce biases yet homogeneous enough to make comparisons meaningful. As we shall see, this sweet spot differs greatly depending on the discipline and the aim of the evidence synthesis.

3.1. Measuring heterogeneity

As shown in table 1, important steps in a meta-analysis include assigning a weight to each study, based on its quality, and estimating the *overall effect* of all the primary studies taken together. Cases with low heterogeneity, that is, with little variation between primary studies, are deemed "simple." Here, any differences in the observed

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

Table 1. Key Terms

Term	Definition	Comments	References
Meta-research (meta-science)	The study of the process of research itself	The aim of meta-research is to evaluate and improve research practices.	(Ioannidis 2018; Nakagawa et al. 2020)
Research synthesis	The process of bringing together findings and attributes from different publications	This is meant to be the umbrella term that encompasses the various types of synthesis outlined next.	(Koricheva and Gurevitch 2013; Nakagawa et al. 2020)
Evidence synthesis	Quantitative or qualitative aggregation of results from primary studies	In the biological literature, this term refers primarily to systematic reviews and meta-analyses but can include other types of reviews, such as systematic maps and rapid reviews. This is based on the old Cochrane Library terminology (from 2003), which has been eclipsed in more recent *Cochrane Handbooks*, where the focus goes directly to systematic review. Here, *systematic review* has become the more general term, referring to various types of synthesis. I will follow the biological literature and use the term to encompass systematic review and meta-analysis.	(Higgins, Thomas, et al. 2019; Nakagawa et al. 2020)
E-synthesis	A Bayesian framework for analyzing and interpreting causal indicators in pharmacology	This is a Bayesian approach to aggregating results from primary studies, distinct from and complementary to evidence synthesis.	(De Pretis et al. 2019; Osimani 2020)
Systematic review	An amalgamation of all empirical evidence that fits prespecified eligibility criteria in order to answer a specific research question	Systematic reviews use explicit and systematic methods for amalgamating data that are aimed at reducing bias. Specifically, 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. In medicine, *systematic review* is sometimes used as the more general term, encompassing various types of synthesis.	(Foo et al. 2021; Higgins, Thomas, et al. 2019; Nakagawa et al. 2020; Siddaway et al. 2019)

(*Continued*)

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

Table 1. (*Continued*)

Term	Definition	Comments	References
Meta-analysis	A statistical tool for (i) determining the overall effect of an intervention or phenomenon, (ii) arbitrating between contradictory results, and (iii) generating generalizations and investigating their scope	A meta-analysis includes all the steps of a systematic review; thus, it is sometimes referred to as a *systematic review with qualitative synthesis*. Main steps: • Conduct systematic review. • Extract data from primary research. • Calculate *effect size* for each primary study. • Assign weight to each study. • Conduct *risk-of-bias assessment.* • Conduct sensitivity analysis (how much *heterogeneity* is explained by known factors?). • Determine overall effect.	(Deeks et al. 2019; Gurevitch et al. 2018; Nakagawa and Cuthill 2007)
Effect size	A statistical parameter that can be used to compare the results of different studies in which a common effect of interest has been measured	The standardized effect size is a way to transform the data on the results of each study into a standardized parameter that can be analyzed 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 nonsignificant results will have small effect sizes, whereas those that have negative results will have negative effect sizes. A study finding that the total biomass of plants increased with the introduction of herbivores would have a negative effect size.	(Deeks et al. 2019; Koricheva et al. 2013; Nakagawa and Cuthill 2007)
Risk-of-bias assessment	A tool comprising a questionnaire with an associated algorithm that estimates the likelihood (e.g., low, some concern, high) of various biases in a synthesis	There are a number of different risk-of-bias tools available, which are tailored to different types of studies. Examples of bias in primary studies include confounding biases, exposure selection biases, misclassified comparison biases, performance biases, detection biases, and outcome reporting and outcome assessment biases. Examples of biases in secondary research include searching, screening, data extraction, and data synthesis biases.	(Boutron et al. 2019; Higgins, Savović, et al. 2019; Konno et al. 2024; Sterne et al. 2019) A comprehensive list of biases can be found at catalogofbias.org, a platform developed by the Centre for Evidence-Based Medicine (CEBM) at Oxford University.
Heterogeneity	The presence of variation in true effect sizes underlying different studies	Heterogeneity or "true variance" is understood as the variation in effect sizes that remains after we have accounted for sampling error. Typical measures of heterogeneity include I^2, Cochran's Q, and τ^2.	(Higgins 2008; Higgins and Thompson 2002; Senior et al. 2016)

effects between primary studies are assumed to be due to sampling error (Senior et al. 2016). 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—that is, those with high heterogeneity between primary studies or nonindependent[1] data sets—researchers use *random effects* or *multilevel* statistical models (Nakagawa and Santos 2012). In these cases, the level of heterogeneity (or nonindependence) affects the weighting of the primary studies: rather than 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, a higher sample size only protects against some types of bias, but not all, so our confidence in the overall effect size should not be inflated just because of large 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 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 aimed at identifying what, if any, biases can be identified in the literature as a whole and the meta-analysis in particular (Higgins et al., 2019; Sterne et al. 2019). 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 nonindependence are much higher than those in medicine; Konno et al. 2024). These statistical tests offer a standardized method for interpreting 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 and Thompson 2002). In fact, the widespread use of I^2 has allowed for the adoption of heterogeneity benchmarks, with 25%, 50%, and 75%, respectively, referring to small, medium, and high heterogeneity (Senior et al. 2016).

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 (i.e., in 30%–55% of studies, there are differences in effect sizes that cannot be explained by sampling error; Cuijpers et al. 2021; Higgins and 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. Here, heterogeneity comes from differences in the population samples (e.g., age group, geographical region, gender), experimental setup, or intervention procedures (e.g.,

[1] *Nonindependence* 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 (i) when multiple proxies are used to measure a certain trait (e.g., mating success, breeding success, and survival as a proxy for fitness) or (ii) when studies span multiple species and there is phylogenetic relatedness between a subset of these species.

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

different dosages). In biology, heterogeneity is typically much higher (Nakagawa and Santos 2012; Whittaker 2010). More specifically, Senior et al. (2016) show that ecologists should expect an I^2 of 90% or more, with only 4.65% of studies having an I^2 value of 0. This is not surprising because primary research in biology can vary in many more ways, including the 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. 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).

3.2. The problem of heterogeneity

The worry with heterogeneity in evidence synthesis is that too many differences between primary studies will render comparisons between them difficult, misleading, or even completely meaningless (Carpenter 2020; Ioannidis et al. 2007; Spake et al. 2022; Whittaker 2010). 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 in each study is different, and there is no way to calculate an overall effect. More specifically, heterogeneity between primary studies can create artificial differences between effect sizes of different studies, 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 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 it also created side effects in the patients, then pooling the studies could obscure the percentage of patients experiencing adverse effects, effectively concealing the problem.

Similar arguments can be made for heterogeneity in biology, where heterogeneity is much larger (see sec. 3.1; 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 3.1, 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 because the former does not require the uprooting of the entire individual—something that cannot realistically be performed on trees and only works 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 artifact 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, he argues that 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

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

are kept constant could reveal that high levels of species richness matter more for productivity in forests than they do 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 because he believes biologists simply do not have the right kind of data to conduct meta-analyses of sufficiently high quality.

Admittedly, Whittaker's article is, by now, 15 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 thoroughly scrutinized (Gurevitch et al. 2018; Koricheva and Gurevitch 2014; Nakagawa et al. 2017; Nakagawa and Santos 2012). Biologists currently have more sophisticated statistical tools at their disposal (Nakagawa et al. 2022; Nakagawa and Santos 2012), as well as more collective experience in conducting meta-analyses and overcoming the various problems that arise (Nakagawa and Cuthill 2007; Sánchez-Tójar et al. 2018, 2020). Still, the high levels of heterogeneity worry even the staunchest advocates of meta-analysis in biology.

There are two additional worries expressed in the biological literature. The first is that heterogeneity is not adequately reported in biological meta-analyses (Nakagawa and Santos 2012; O'Connor et al. 2017; Schielzeth and 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—that is, if all studies showed no (or at least nonsignificant) 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 small 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 generalizations (Nakagawa and Cuthill 2007; Spake et al. 2022). For example, Nakagawa and 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 centered 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. 594–95). The worry seems to be that only when all the primary studies in a meta-analysis focus on the same type of experiment or species are claims about that experiment or species legitimate. In contrast, when the meta-analysis includes data from multiple species, experimental setups, and so forth, the overall effect size might not be equally representative of or 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

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

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.

With the exposition of heterogeneity and its main critiques in place, it is now time to take a closer look at the motivations for engaging in evidence synthesis.

4. 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 and Gurevitch 2013; 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. Before delving in, I should note that these different goals are not mutually exclusive. The same tool can be used for different goals and can also be used for more than one goal simultaneously. Thus, for example, the same meta-analysis can be used to generate causal confidence and also make sense of any contradictory results. However, it is important, in order to avoid confusion and opaqueness, for the goals of the meta-analysis to be clearly stated and distinguished.

4.1. 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, such as randomized controlled trials (RCTs), is necessarily limited in terms of sample size. In RCTs, patients are randomized and placed in the treatment or control groups, the latter of which receive a placebo rather than the treatment. RCTs are viewed positively (compared with, say, observational studies) because they aim to control for confounding variables and identify genuine causal links between the intervention and the outcome. However, they are usually quite small because of various inherent difficulties: acquiring subjects, testing rare conditions, availability of drugs, costs of conducting trials, 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 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 and 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 treat 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 some evidence of a causal relationship between two variables, yet our confidence regarding each individual study is usually small. A meta-analysis showing

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

that many studies identify the same causal relationship will increase our confidence that the causal relationship truly holds. 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 overall effect than a meta-analysis of drug B, then drug A is more effective for the treatment of depression.

One interpretation of this use of meta-analysis is that by increasing our confidence that the results of clinical trials have indeed established causal links between certain interventions and outcomes, meta-analysis improves the quality of primary research. In this view, RCTs are referred to as the "gold standard" of evidence, making systematic reviews and meta-analyses, which attempt to eliminate or at least minimize their deficits, the "platinum standard" of evidence (see discussion in Stegenga 2011). This is quite a controversial interpretation of the use of evidence synthesis, which has generated a lot of critique (Ioannidis 2016; Stegenga 2011; Worrall 2002). Moreover, as we shall see in section 5, this is the context in which heterogeneity is most problematic.

4.2. Arbitrating between contradictory results

Sometimes, the results of primary studies do not merely differ but are 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 sec. 3.1). For example, a meta-analysis could reveal that the studies showing no effects of drug A have extremely small sample sizes and should be weighted less heavily. Thus, a meta-analysis can show that some apparent contradictions between primary studies can 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 and Poulin 2012). Vote counting involves sorting primary research into three categories (significant results in favor of hypothesis, significant results against hypothesis, and nonsignificant results), determining which category has the highest number of studies, and declaring that category the "winner" (Koricheva and Gurevitch 2013). A major problem with vote counting is 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 and Gurevitch 2013; Nakagawa et al. 2017; Nakagawa and Poulin 2012).

Meta-analyses can also reveal how different measurements of a certain phenomenon can lead to different conclusions and provide information about how to deal with the resulting contradictory conclusions. Consider the case of biodiversity trends—that is, whether biodiversity is increasing or decreasing over the last decades. While many studies have concluded that biodiversity is decreasing, there have been some studies that demonstrate an increase in biodiversity. This is interesting but also potentially problematic; it can have an effect on conservation policy and funding allocation because it can be used as "evidence" for decreasing the

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

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 that found no changes or found 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 the conservation of native species in Europe, we can discount the studies that measure species turnover (see also sec. 6).

4.3. Exploring the scope of generalizations

Perhaps the least well-known but, in my opinion, the most useful goal of meta-analysis is as a tool for testing the scope of generalizations.[2] At first glance, it seems similar to the goal outlined in the previous section because it is also 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. However, the situation becomes trickier when scientists try to determine exactly how enemy release manifests in each case and what conclusions can be drawn from it (Heger and 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 a number of cases where alien plants attracted native herbivores, and these herbivores had a significant negative effect on seed production and plant survival (both of which are important for a successful invasion; Maron and Vilà 2001). Studies at different scales also tend to yield contradictory results because larger-scale biogeographical analyses primarily show a reduction in the diversity of enemies in the introduced range compared with the native range, whereas 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) analyzed 63 manipulative field studies of plant invasions that included the effect of herbivores on the outcome of the invasion (i.e., they included primary studies where herbivores facilitated the plant invasion and where they hindered it). At first glance, it seemed that there was stronger evidence against the enemy release hypothesis: there were cases where native herbivores decreased the abundance of alien plants (i.e., plants encountered new enemies) and cases where alien herbivores (their existing enemies) increased the abundance of alien plants. However, they also found that the negative effect of native

[2] The ideas in this section, along with the following two sections, were first discussed by Elliott-Graves (2023, sec. 4.2.2). What follows is a more in-depth investigation of these issues with a special focus on the role of heterogeneity.

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

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 realized 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 where 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, which 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, which 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 generalizations 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 2023; Doak et al. 2008). This means that ecological generalizations are often limited in scope (Elliott-Graves 2023; Mitchell 2000). This creates problems for ecological research because generalizations form the basis for some types of explanations and most predictions; a generalization breaking down translates into knowledge not being transferable across systems or across time periods (Catford et al. 2022; Spake et al. 2023). While ecologists are 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 in a system (Catford et al. 2022; Doak et al. 2008; Mouquet et al. 2015). Thus, any information on the scope and limits of a generalization can be incredibly useful; it can make the difference between a successful and unsuccessful intervention. In the case of enemy release, knowing that the enemy release effect is overshadowed by generalist herbivores can have important effects on policy. Here, 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 nonnative 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 two sections, I will examine the cases where heterogeneity is genuinely problematic and when it is not.

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

5. 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 it. The problem is that unreported heterogeneity implies that there is no significant heterogeneity in the studies, so any overall effect is taken at face value. If, however, there is significant heterogeneity in the effect sizes, then the issues outlined in section 3 hold: We cannot be sure that the overall effect 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 4.1—namely, "generating causal confidence." Recall that this use of meta-analysis involves pooling results from different studies in the hope of generating a result with higher statistical significance, thus increasing our confidence in the result. Here, researchers treat the primary studies as though they are replicates of each other; that is, they assume that there is a high level of homogeneity between the studies, and thus 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 3, where the meta-analysis is aiming to show a significant effect size for a certain drug, yet primary studies differ in terms of the dosage administered. In this example, only the higher dose of the drug is actually effective, yet it 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 and further investigated.

6. 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. 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 and Thompson 2002; Nakagawa et al. 2017; Schielzeth and Nakagawa 2022; Senior et al. 2016; see also discussion in sec. 4). When heterogeneity is expected and adequately reported, then researchers have access to numerous methods for further

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

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 analyze 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 multispecies 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.[3]

But how exactly can analyzing heterogeneity be useful? Unlike the context of generating causal confidence, when we are using meta-analyses to arbitrate between contradictory results (sec. 4.2) or examine the scope of generalizations (sec. 4.3), heterogeneity can provide us with valuable information. Starting with the case of contradictory results, heterogeneity is useful when groups of primary studies emerge that display intragroup homogeneity and intergroup heterogeneity, in other words, when heterogeneity clusters in interesting ways. As shown in section 4.2, 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 that showed decreases in biodiversity were those that measured richness or abundance, whereas those that showed no changes or increases in biodiversity were those that measured species turnover. These heterogeneous clusters are quite informative when we are trying to make sense of contradictory results. In this case, they show us that the 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 by 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 that could account for the different results.[4] In the case of biology, if the clusters correlate to particular species, 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 issue in the next paragraph. The point here is that heterogeneous clusters, if properly investigated, can account for contradictory results and can sometimes provide additional information that explains the underlying causes of the contradiction.

The case for preserving and analyzing heterogeneity is even stronger in the context of exploring the scope of generalizations because it is the existence of

[3] 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 or adapted these packages for biological data are also extremely helpful, willing to answer questions and troubleshoot the implementation of the software.

[4] However, as we shall see presently, when we expect low heterogeneity, certain ways of investigating heterogeneity can be misleading because they can create artificial effects.

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

heterogeneity itself that predicts the limits of a generalization and, in some cases, can explain the limits of the generalization in question. I will return to the case of the enemy release hypothesis, outlined in section 4.3. Here, the scientists were able to explain the reason why the primary studies examining the enemy release hypothesis yielded contradictory results because they realized 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 the scope of the enemy release hypothesis—that is, where the mechanism of enemy release was effective and where it was not. In fact, if the researchers had reduced the heterogeneity of their sample in the traditionally approved way—that is, 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 realized 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 vertebrate and invertebrate). If they had excluded one group by default, they would not have realized the limit in scope of the enemy release mechanism (i.e., when it was overshadowed by other mechanisms). Second, failing to understand this would also have prevented the scientists from gaining another insight into biological invasions—namely, that this explains the hitherto perplexing phenomenon that it is much more common for European plants to invade areas outside Europe rather than vice versa. The insight 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 the exotic plants with which they have co-evolved.

In short, "correcting" for heterogeneity—that is, leaving out the primary studies that increase the heterogeneity of the overall effect—can sometimes create more problems than it solves. Here, heterogeneity is a feature rather than a bug, and although 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 the discussion in this section pertains to disciplines such as biology, where high heterogeneity is expected. Moreover, in section 5, 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? It seems that even investigating heterogeneity in medicine is sometimes problematic.

Some medical meta-analyses investigate heterogeneity in the form of *subgroup analyses* (Cook et al. 2004; Cuijpers et al. 2021). Here, studies are divided into two or more subgroups to test whether the pooled effect sizes found in these subgroups differ significantly from each other. Thus, for example, a meta-analysis can divide the pooled individuals into subgroups based on gender, age, geographical region, dosage, or type of environment (e.g., a nursing home) to explore whether there are differences between the groups. At first glance, this seems to be very similar to what Parker et al. (2006) did in the enemy release case. However, in the medical context, subgroup analysis seems to be more dangerous. The problem is that with small

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

sample sizes and low to middling levels of expected heterogeneity, the subgroup analysis could identify effects that are not really there. Thus, instead of identifying genuine causes of heterogeneity, it can create the illusion of genuine differences (Cook et al. 2004; Cuijpers et al. 2021; Higgins et al., 2019). Does this undermine the case for heterogeneity?

A noteworthy point to consider is that in discussing the use of subgroup analyses in medicine, many authors argue that in order to avoid the previously described problem, the analysis should only include subgroups that have genuine underpinnings (which are referred to as *biological mechanisms*; Cook, et al. 2004). This can seem somewhat arbitrary. Why are biological differences more important than environmental differences? I think that the answer again lies in expectations surrounding heterogeneity. Humans are all the same species, and in clinical trials, they are treated in quite similar ways. It is therefore reasonable that any differences that emerge, especially when dealing with small sample sizes, are artifacts of the statistics. In contrast, in the case of biological meta-analyses, we are often dealing with multiple species and/or quite different experimental setups, hence the expectations of heterogeneity being around 90%. In the latter case, there *are many known differences* between subgroups (invertebrates vs. vertebrates, generalists vs. specialists); the question is whether these differences are relevant to the question being studied. Thus, it seems that in order for the statistical investigation of heterogeneity to be valuable, prior knowledge of heterogeneity should already have been established.

Still, this does not entirely preclude heterogeneity being useful in medical contexts. It simply shows that we should be careful even when investigating heterogeneity and avoid creating the illusion of heterogeneity through statistical error. While this point is acknowledged in the literature, I think its importance is undervalued. The main focus in the medical literature still seems to be on cautioning against the dangers of subgroup analysis (Cook et al. 2004; Cuijpers et al. 2021; Higgins, Thomas, et al. 2019). While these cautions are valuable, they can be complemented with a better understanding of the potential value of heterogeneity.

7. Conclusion

Heterogeneity is a delicate topic in evidence synthesis and is the subject of significant controversy. I have argued that a blanket approach to heterogeneity is unlikely to be useful. What determines whether heterogeneity is a problem depends on our attitude toward it. If we assume that it is nonexistent, when it does exist, then our synthesis will suffer. If, on the other hand, we have good reasons to expect heterogeneity to exist and explicitly analyze it, then we can end up with a lot more information than we would from an entirely homogeneous set of primary research (Higgins and 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 makes exploring heterogeneity an integral but also useful dimension of evidence synthesis.

Acknowledgments. I would like to thank the biologists and philosophers of the Collaborative Research Center initiative "A Novel Synthesis of Individualisation across Behaviour, Ecology and Evolution: Niche Choice, Niche Conformance, Niche Construction (NC^3)," especially Alfredo Sánchez-Tójar and the participants of the Evidence Synthesis Toolkit Workshop, which took place in Bielefeld (October 2024), for

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

their help in discussing and refining the ideas in this article. I would also like to thank the PhilLiSci group at Bielefeld and two anonymous referees for helpful and constructive feedback on previous versions of the article. The research for this article was supported by the German Research Foundation (DFG) as part of project D02, "Individual-Based Research: Concepts, Epistemology and Integration" (project number 396781820), in the TRR-CRC 212 "A Novel Synthesis of Individualisation across Behaviour, Ecology and Evolution: Niche Choice, Niche Conformance, Niche Construction (NC^3)" (project number 316099922).

References

Beckage, Brian, Louis J. Gross, and Stuart Kauffman. 2011. "The Limits to Prediction in Ecological Systems." *Ecosphere* 2 (11):1–12. https://doi.org/10.1890/ES11-00211.1.

Berchialla, Paola, Daniele Chiffi, Giovanni Valente, and Ari Voutilainen. 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, Jesse A., and Robert M. Golub. 2014. "Meta-Analysis as Evidence: Building a Better Pyramid." *JAMA* 312 (6):603–6. https://doi.org/10.1001/jama.2014.8167.

Boutron, Isabelle, Matthew J. Page, Julian P. T. Higgins, Douglas G. Altman, Andreas Lundh, and Asbjørn Hróbjartsson. 2019. "Considering Bias and Conflicts of Interest among the Included Studies." In *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed., edited by Julian P. T. Higgins, James Thomas, Jacqueline Chandler, et al., chap. 7. Chichester, UK: John Wiley & Sons. https://doi.org/10.1002/9781119536604.ch7.

Bruner, Justin P., and Bennett Holman. 2019. "Self-Correction in Science: Meta-Analysis, Bias and Social Structure." *Studies in History and Philosophy of Science Part A* 78 (December):93–97. https://doi.org/10.1016/j.shpsa.2019.02.001.

Cadotte, Marc W., Lea R. Mehrkens, and Duncan N. L. Menge. 2012. "Gauging the Impact of Meta-Analysis on Ecology." *Evolutionary Ecology* 26 (5):1153–67. https://doi.org/10.1007/s10682-012-9585-z.

Carpenter, Christopher J. 2020. "Meta-Analyzing Apples and Oranges: How to Make Applesauce Instead of Fruit Salad." *Human Communication Research* 46 (2–3):322–33. https://doi.org/10.1093/hcr/hqz018.

Catford, Jane A., John R. U. Wilson, Petr Pyšek, Philip E. Hulme, and Richard P. Duncan. 2022. "Addressing Context Dependence in Ecology." *Trends in Ecology & Evolution* 37 (2):158–70. https://doi.org/10.1016/j.tree.2021.09.007.

Chen, Haoyu, and Vishal Jhanji. 2012. "Survey of Systematic Reviews and Meta-Analyses Published in Ophthalmology." *British Journal of Ophthalmology* 96 (March):896–99. https://doi.org/10.1136/bjophthalmol-2012-301589.

Colautti, Robert I., Anthony Ricciardi, Igor Grigorovich, and Hugh J. MacIsaac. 2004. "Is Invasion Success Explained by the Enemy Release Hypothesis?" *Oikos* 7 (8):721–33.

Cook, D.I., Gebski, V.J. and Keech, A.C. (2004), Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180:289–291. https://doi.org/10.5694/j.1326-5377.2004.tb05928.x

Cook, David I., Val J. Gebski, and Anthony C. Keech. 2004. "Subgroup Analysis in Clinical Trials." *Medical Journal of Australia* 180 (6):289–91. https://doi.org/10.5694/j.1326-5377.2004.tb05928.x.

Cuijpers, Pim, Jason W. Griffin, and Toshi A. Furukawa. 2021. "The Lack of Statistical Power of Subgroup Analyses in Meta-Analyses: A Cautionary Note." *Epidemiology and Psychiatric Sciences* 30 (December):e78. https://doi.org/10.1017/S2045796021000664.

Deeks, Jonathan J., Julian P. T. Higgins, and Douglas G. Altman, on behalf of the Cochrane Statistical Methods Group. 2019. "Analysing Data and Undertaking Meta-Analyses." In *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed., edited by Julian P. T. Higgins, James Thomas, Jacqueline Chandler, et al., chap. 10. Chichester, UK: John Wiley & Sons. https://doi.org/10.1002/9781119536604.ch10.

Dettori, Joseph R., Daniel C. Norvell, and Jens R. Chapman. 2022. "Fixed-Effect vs Random-Effects Models for Meta-Analysis: 3 Points to Consider." *Global Spine Journal* 12 (7):1624–26. https://doi.org/10.1177/21925682221110527.

Doak, Daniel F., James A. Estes, Benjamin S. Halpern, et al. 2008. "Understanding and Predicting Ecological Dynamics: Are Major Surprises Inevitable?" *Ecology* 89 (4):952–61. https://doi.org/10.1890/07-0965.1.

Egger, Matthias, Shah Ebrahim, and George Davey Smith. 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: Cambridge University Press.

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

Fieseler, Clare. 2021. "The Case against the Concept of Biodiversity." *Vox.* https://www.vox.com/22584103/biodiversity-species-conservation-debate.

Fletcher, Samuel C. 2022. "Replication Is for Meta-Analysis." *Philosophy of Science* 89 (5):960–69. https://doi.org/10.1017/psa.2022.38.

Fontelo, Paul, and Fang Liu. 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, Yong Zhi, Rose E. O'Dea, Julia Koricheva, Shinichi Nakagawa, and Malgorzata Lagisz. 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):1705–20. https://doi.org/10.1111/2041-210X.13654.

Gurevitch, Jessica, Julia Koricheva, Shinichi Nakagawa, and Gavin Stewart. 2018. "Meta-Analysis and the Science of Research Synthesis." *Nature* 555 (7695):175–82. https://doi.org/10.1038/nature25753.

Hardwicke, Tom E., Stylianos Serghiou, Perrine Janiaud, et al. 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, Tina, and Jonathan Jeschke. 2014. "The Enemy Release Hypothesis as a Hierarchy of Hypotheses." *Oikos* 123 (6):741–50. https://doi.org/10.1111/j.1600-0706.2013.01263.x.

Higgins, Julian P. T. 2008. "Commentary: Heterogeneity in Meta-Analysis Should Be Expected and Appropriately Quantified." *International Journal of Epidemiology* 37 (5):1158–60. https://doi.org/10.1093/ije/dyn204.

Higgins, Julian P. T., Jelena Savović, Matthew J. Page, Roy G. Elbers, and Jonathan A. C. Sterne. 2019. "Assessing Risk of Bias in a Randomized Trial." In *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed., edited by Julian P. T. Higgins, James Thomas, Jacqueline Chandler, et al., chap. 8. Chichester, UK: John Wiley & Sons. https://doi.org/10.1002/9781119536604.ch8.

Higgins, Julian P. T., James Thomas, Jacqueline Chandler, et al., eds. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Chichester, UK: John Wiley & Sons. https://doi.org/10.1002/9781119536604.fmatter.

Higgins, Julian P. T., and S. G. Thompson. 2002. "Quantifying Heterogeneity in a Meta-Analysis." *Statistics in Medicine* 21 (11):1539–58. https://doi.org/10.1002/sim.1186.

Holman, Bennett. 2019. "In Defense of Meta-Analysis." *Synthese* 196 (8):3189–211. https://doi.org/10.1007/s11229-018-1690-2.

Houlahan, Jeff, Shawn McKinney, Michael Anderson, and Brian McGill. 2017. "The Priority of Prediction in Ecological Understanding." *Oikos* 126 (1):1–7. https://doi.org/10.1111/oik.03726.

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

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

Ioannidis, John P. A. 2018. "Meta-Research: Why Research on Research Matters." *PLOS Biology* 16 (3):e2005468. https://doi.org/10.1371/journal.pbio.2005468.

Ioannidis, John P. A., Nikolaos A. Patsopoulos, and Evangelos Evangelou. 2007. "Uncertainty in Heterogeneity Estimates in Meta-Analyses." *BMJ* 335 (7626):914–16. https://doi.org/10.1136/bmj.39343.408449.80.

Jukola, Saana. 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 (April):35–41. https://doi.org/10.1016/j.shpsc.2017.02.001.

Kaunisto, Sirpa, Laura V. Ferguson, and Brent J. Sinclair. 2016. "Can We Predict the Effects of Multiple Stressors on Insects in a Changing Climate?" *Current Opinion in Insect Science* 17 (October):55–61. https://doi.org/10.1016/j.cois.2016.07.001.

Konno, Ko, James Gibbons, Ruth Lewis, and Andrew S. Pullin. 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, Julia, and Jessica Gurevitch. 2013. "Place of Meta-Analysis among Other Methods of Research Synthesis." In *Handbook of Meta-Analysis in Ecology and Evolution*, edited by Julia Koricheva, Jessica Gurevitch, and Kerrie Mengersen, 3–13. Princeton, NJ: Princeton University Press.

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

Koricheva, Julia, and Jessica Gurevitch. 2014. "Uses and Misuses of Meta-Analysis in Plant Ecology." *Journal of Ecology* 102 (4):828–44. https://doi.org/10.1111/1365-2745.12224.

Koricheva, Julia, Jessica Gurevitch, and Kerrie Mengersen, eds. 2013. *Handbook of Meta-Analysis in Ecology and Evolution*. Princeton, NJ: Princeton University Press.

Kovaka, Karen. 2022. "Meta-Analysis and Conservation Science." *Philosophy of Science* 89 (5):980–90. https://doi.org/10.1017/psa.2022.68.

LaCaze, Adam, and Barbara Osimani, eds. 2020. *Uncertainty in Pharmacology: Epistemology, Methods, and Decisions*. Boston Studies in the Philosophy and History of Science. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-29179-2.

Lawton, J. H. 1999. "Are There General Laws in Ecology?" *Oikos* 84 (2):177–92. https://doi.org/10.2307/3546712.

Maron, John L., and Montserrat Vilà. 2001. "When Do Herbivores Affect Plant Invasion? Evidence for the Natural Enemies and Biotic Resistance Hypotheses." *Oikos* 95 (3):361–73. https://doi.org/10.1034/j.1600-0706.2001.950301.x.

Marsden, Emma, Kara Morgan-Short, Sophie Thompson, and David Abugaber. 2018. "Replication in Second Language Research: Narrative and Systematic Reviews and Recommendations for the Field." *Language Learning* 68 (2):321–91. https://doi.org/10.1111/lang.12286.

Maziarz, Mariusz. 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 (February):159–67. https://doi.org/10.1016/j.shpsa.2021.11.007.

Mitchell, Sandra D. 2000. "Dimensions of Scientific Law." *Philosophy of Science* 67 (2):242–65. https://doi.org/10.2307/188723?refreqid=search-gateway:148949eeb4f88afdc4c4ac3f69c73275.

Mitchell, Sandra D. 2002. "Ceteris Paribus—an Inadequate Representation for Biological Contingency." *Erkenntnis* 57 (3):329–50.

Mouquet, Nicolas, Yvan Lagadeuc, Vincent Devictor, et al. 2015. "Predictive Ecology in a Changing World." *Journal of Applied Ecology* 52 (5):1293–310. https://doi.org/10.1111/1365-2664.12482.

Mueller-Langer, Frank, Benedikt Fecher, Dietmar Harhoff, and Gert G. Wagner. 2019. "Replication Studies in Economics—How Many and Which Papers Are Chosen for Replication, and Why?" *Research Policy* 48 (1):62–83.

Nakagawa, Shinichi, and Innes C. Cuthill. 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, Shinichi, Julia Koricheva, Malcolm Macleod, and Wolfgang Viechtbauer. 2020. "Introducing Our Series: Research Synthesis and Meta-Research in Biology." *BMC Biology* 18 (1):20. https://doi.org/10.1186/s12915-020-0755-0.

Nakagawa, Shinichi, Malgorzata Lagisz, Michael D. Jennions, et al. 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, Shinichi, Daniel W. A. Noble, Alistair M. Senior, and Malgorzata Lagisz. 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, Shinichi, and Robert Poulin. 2012. "Meta-Analytic Insights into Evolutionary Ecology: An Introduction and Synthesis." *Evolutionary Ecology* 26 (5):1085–99. https://doi.org/10.1007/s10682-012-9593-z.

Nakagawa, Shinichi, and Eduardo S. A. Santos. 2012. "Methodological Issues and Advances in Biological Meta-Analysis." *Evolutionary Ecology* 26 (5):1253–74. https://doi.org/10.1007/s10682-012-9555-5.

O'Connor, Mary I., Andrew Gonzalez, Jarrett E. K. Byrnes, et al. 2017. "A General Biodiversity–Function Relationship Is Mediated by Trophic Level." *Oikos* 126 (1):18–31. https://doi.org/10.1111/oik.03652.

Osimani, Barbara. 2020. "Epistemic Gains and Epistemic Games: Reliability and Higher Order Evidence in Medicine and Pharmacology." In *Uncertainty in Pharmacology*, vol. 338, edited by Adam LaCaze and Barbara Osimani, 345–72. Boston Studies in the Philosophy and History of Science. Cham, Switzerland: Springer. https://link.springer.com/chapter/10.1007/978-3-030-29179-2_15.

Parker, John D. et al. (2006) Opposing Effects of Native and Exotic Herbivores on Plant Invasions.Science311,1459-1461.DOI:10.1126/science.1121407

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press

Pilotto, Francesca, Ingolf Kühn, Rita Adrian, et al. 2020. "Meta-Analysis of Multidecadal Biodiversity Trends in Europe." *Nature Communications* 11 (1):1. https://doi.org/10.1038/s41467-020-17171-y.

De Pretis, Francesco, Jürgen Landes, and Barbara Osimani. 2019. "E-Synthesis: A Bayesian Framework for Causal Assessment in Pharmacosurveillance." *Frontiers in Pharmacology* 10 (December):1317. https://doi.org/10.3389/fphar.2019.01317.

Pyron, R. Alexander. 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, Jani. 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, Felipe. 2016. "Can the Behavioral Sciences Self-Correct? A Social Epistemic Study." *Studies in History and Philosophy of Science Part A* 60 (December):55–69. https://doi.org/10.1016/j.shpsa.2016.10.002.

Sánchez-Tójar, Alfredo, Nicholas P. Moran, Rose E. O'Dea, Klaus Reinhold, and Shinichi Nakagawa. 2020. "Illustrating the Importance of Meta-Analysing Variances alongside Means in Ecology and Evolution." *Journal of Evolutionary Biology* 33 (9):1216–23. https://doi.org/10.1111/jeb.13661.

Sánchez-Tójar, Alfredo, Shinichi Nakagawa, Moisès Sánchez-Fortún, et al. 2018. "Meta-Analysis Challenges a Textbook Example of Status Signalling and Demonstrates Publication Bias." *eLife* 7 (November):e37385. https://doi.org/10.7554/eLife.37385.

Schielzeth, Holger, and Shinichi Nakagawa. 2022. "Conditional Repeatability and the Variance Explained by Reaction Norm Variation in Random Slope Models." *Methods in Ecology and Evolution* 13 (6):1214–23. https://doi.org/10.1111/2041-210X.13856.

Senior, Alistair M., Catherine E. Grueber, Tsukushi Kamiya, et al. 2016. "Heterogeneity in Ecological and Evolutionary Meta-Analyses: Its Magnitude and Implications." *Ecology* 97 (12):3293–99. https://doi.org/10.1002/ecy.1591.

Siddaway, Andy P., Alex M. Wood, and Larry V. Hedges. 2019. "How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses." *Annual Review of Psychology* 70: 747–70. https://doi.org/10.1146/annurev-psych-010418-102803.

Spake, Rebecca, Diana E. Bowler, Corey T. Callaghan, et al. 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, Rebecca, Rose E. O'Dea, Shinichi Nakagawa, et al. 2022. "Improving Quantitative Synthesis to Achieve Generality in Ecology." *Nature Ecology & Evolution* 6 (12):1818–28. https://doi.org/10.1038/s41559-022-01891-z.

Stegenga, Jacob. 2011. "Is Meta-Analysis the Platinum Standard of Evidence?" *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 42 (4):497–507. https://doi.org/10.1016/j.shpsc.2011.07.003.

Sterne, Jonathan A. C., Miguel A. Hernán, Alexandra McAleenan, Barnaby C. Reeves, and Julian P. T. Higgins. 2019. "Assessing Risk of Bias in a Non-Randomized Study." In *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed., edited by Julian P. T. Higgins, James Thomas, Jacqueline Chandler, et al., chap. 25. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119536604.ch25.

Taylor, Amy, and Marcus Munafò. 2016. "Triangulating Meta-Analyses: The Example of the Serotonin Transporter Gene, Stressful Life Events and Major Depression." *BMC Psychology* 4 (1):23. https://doi.org/10.1186/s40359-016-0129-0.

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

Watkins, Hannah V., Helen F. Yan, Jillian C. Dunic, and Isabelle M. Côté. 2021. "Research Biases Create Overrepresented 'Poster Children' of Marine Invasion Ecology." *Conservation Letters* 14 (3):e12802. https://doi.org/10.1111/conl.12802.

Whittaker, Robert J. 2010. "Meta-Analyses and Mega-Mistakes: Calling Time on Meta-Analysis of the Species Richness–Productivity Relationship." *Ecology* 91 (9):2522–33. https://doi.org/10.1890/08-0968.1.

Worrall, John. 2002. "What Evidence in Evidence-Based Medicine?" *Philosophy of Science* 69 (S3):S316–30. https://doi.org/10.1086/341855

Cite this article: Elliott-Graves, Alkistis. 2026. "Reevaluating Heterogeneity in Evidence Synthesis." *Philosophy of Science* 93 (3):541–560. https://doi.org/10.1017/psa.2025.10176

https://doi.org/10.1017/psa.2025.10176 Published online by Cambridge University Press