Meta-analytic approaches and ‘new’ effect sizes to account for heterogeneous effects in comparative physiology

Daniel W.A. Noble, Patrice Potter, Malgorzata Lagisz, Sammy Burke, Szymek Drobniak, Rose O’Dea, Shinichi Nakagawa

Division of Ecology and Evolution, Research School of Biology, The Australian National University, Canberra, ACT 2600

School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW

## Abstract

Meta-analysis is a tool that provides comparative physiologists with powerful, unbiased and quantitatively informed answers to topical research questions by synthesising effect sizes across disparate studies. Meta-analysis’ core objective and strength is to distill published research results into standardised effect sizes that can be weighted (i.e., by their inverse sampling variance), and which can be compared across broad sets of research designs, study systems and species. Estimating overall effect sizes, and understanding what drives effect variability, provides opportunities to model how organisms will respond to pressing global challenges. Despite this ambition, research designs in comparative physiology can appear, at the outset, as being vastly different to each other (e.g., using different temperatures or treatment dosages). Differences in treatments across studies has led many to believe that meta-analysis is an exercise in comparing “apples with oranges”. Here, we dispel this myth by showing how standardised effects sizes can be used in conjunction with powerful multi-level meta-regression models to both account for factors driving differences across studies and make them more comparable. In addition, we formalise effect size measures that provide comparative physiologists with a means to directly make effect measures comparable, removing ‘nuisance heterogeneity’ without the need to resort to more complex statistical models that may be harder to interpret. We also describe more general approaches that can be usefully applied to a variety of different contexts to derive ‘new’ effects sizes and sampling variances. The effect sizes and their corresponding sampling variances that we describe should help physiologists deal with common treatment differences across studies. By using effect sizes that account for components of effect heterogeneity, in combination with existing meta-analytic models, comparative physiologists can explore exciting new questions by making results from large-scale data sets collected from the literature more accessible and widely interpretable.

## Introduction

A meta-analysis provides quantitatively informed answers to research questions by synthesising research results across studies and identifying sources of variation (Arnqvist and Wooster, 1995; Borenstein, 2019; Cooper et al., 2009; Gurevitch and Hedges, 1999; Gurevitch et al., 2018; Koricheva et al., 2013; Nakagawa et al., 2017). Meta-analysing independent studies provides greater statistical power and precision than what any individual study on its own would be able to provide on a question - particularly given that most empirical studies are already under-powered in many areas of biology (Button et al., 2013; Forstmeier et al., 2017; Jennions and Møller, 2003). By expressing study effects on a common scale (i.e., standardised effect size) a meta-analyst can gain broader insight into the direction and efficacy of a particular treatment effect or the strength of a relationship between two variables of interest (Gurevitch et al., 2018; Koricheva et al., 2013). Quantitative syntheses, including meta-analyses, have provided comparative physiologists with powerful insights on pressing global challenges - from testing whether physiological plasticity can buffer organisms against climate change (e.g, Seebacher et al., 2015) to the degree to which endocrine disrupting chemicals, such as bis-phenol A (BPA), impact aquatic organisms (e.g., Wu and Seebacher, 2020).

Meta-analysis is a well-developed research field with a long history in statistics, medicine, psychology and ecology (Glass, 2015; Gurevitch et al., 2018; Nakagawa et al., 2017). It has emerged as the ‘gold-standard’ across various disciplines for quantitative synthesis and is becoming increasingly more prominent in the field of comparative physiology. Formal meta-analyses are part of a systematic review and involve a systematic literature-based search, extraction of data from studies meeting specific inclusion criteria, and the subsequent calculation of effects sizes from summary statistics provided by papers. A meta-analysis should follow and report upon this entire process carefully to ensure the review process and data collation, are as transparent and reproducible as possible (i.e., see the Preferred Reporting Items for Systematic Reviews and Meta-Analyses in Ecology and Evolution (PRISMA-EcoEvo) protocol, O’Dea et al., 2021). Aside from providing an up-to-date overview of a research field there are three major goals of meta-analysis: 1) to provide an overall mean estimate of a treatment effect or relationship, 2) to quantify effect size variance and understand key drivers explaining differences in effects across studies and 3) attempt to identify gaps and publication biases (Borenstein, 2019; Cooper et al., 2009; Gurevitch and Hedges, 1999; Gurevitch et al., 2018; Koricheva et al., 2013; Nakagawa et al., 2017). A critically important feature of a meta-analysis is its ability to account for sampling error (Koricheva et al., 2013). The inverse of sampling error variance is used to weight studies in the analysis and make overall estimates more precise. While many meta-analyses might employ unweighted analyses, these studies will not gain greater precision on estimates because sampling variance cannot be formally accounted for in the analysis, even if the overall mean itself is unbiased (Morrissey, 2016).

Despite its widespread adoption and well established methodological procedures, meta-analysis is often criticised for mixing ‘apples and oranges’ – in other words, mixing effects from studies that lack comparability (Arnqvist and Wooster, 1995; Carpenter, 2020; Gallo, 1978; Glass, 2015; Gurevitch et al., 2018; Stewart, 2010). Lack of comparability could result from studies differing in experimental design, species, measurement variables and sampling units, which can impact the generalisability of meta-analytic results (Arnqvist and Wooster, 1995; Stewart, 2010). Indeed, study specific differences in comparative physiology are widespread. For example, experimental designs can vary in the temperatures that they apply or in the dosages of chemical modifiers used in experimental treatments (e.g., Wu and Seebacher, 2020). To the uninitiated, this concern might not be unfounded given that heterogeneity in effects is often high in ecological and evolutionary meta-analyses (Senior et al., 2016), telling us that the effect size varies a great deal from one population to another. However, to many, having highly heterogeneous effects “…is the spice of life” (p.g., 519; Cooper et al., 2019) because it provides opportunities to explore the reasons for why effects vary within and across studies (Borenstein, 2019; Cooper et al., 2019; Glass, 2015; Gurevitch et al., 2018).

The goal of our review is to briefly overview the process of meta-analysis, making a case for why heterogeneity in a meta-analysis is a good thing for answering fundamental research problems and guiding research directions. We introduce the idea of ‘nuisance heterogeneity’, showing comparative physiologists that it can be comfortably dealt with by re-formalising many existing effect sizes measures and/or by using multilevel meta-regression models. Indeed, a combination of these approaches might often be ideal. We first conduct a survey of published meta-analyses in the field of comparative physiology to gauge methodological and meta-analytic practices – determining what types of effect sizes are commonly used, how often effect size estimates are impacted by ‘nuisance heterogeneity’ and whether meta-analytic models (i.e., models that account for sampling variance) are commonly applied. We then formalise alternative effect sizes and their associated sampling variance to provide comparative physiologists opportunities to explicitly incorporate treatment differences into the effect size estimates to ease their interpretation. Finally, we describe how more complex treatment differences, such as dosage differences (which may have non-linear effects) can be accommodated using multilevel meta-regression models. We hope to show how expanding the meta-analytic toolkit for comparative physiologists will provide new opportunities to address global scale problems that attempt to understand how organisms will cope with rapidly changing environments in the future.

## The Apples and Oranges ‘Problem’ in Comparative Physiology

Meta-analysis is always concerned with effect heterogeneity; in other words, the factors that drive differences in the direction and magnitude an of effect within and across studies (Borenstein, 2019; Cooper et al., 2019; Gurevitch et al., 2018; Nakagawa et al., 2017). The concept of heterogeneity is vitally important because it tells us how general a specific conclusion is likely to be, how much of the variance we see is the result of real biology (not just sampling variance), and provides a way to understand what the key biological factors are leading to an effect (Borenstein, 2019; Nakagawa et al., 2017). There are a plethora of phenomena driving effect heterogeneity, some of which are interesting and others that are less so.

Generally, contributors to effect size heterogeneity fall within a number of broad categories. First, sampling variance contributes to heterogeneity; the variability driven by the inherent uncertainty in sampling a given population. We cannot sample the entire population and nor do we necessarily want to. As a consequence, we need to rely on a sub-sample to estimate a population parameter (e.g., correlation between two traits). How many independent samples we collect impacts our precision in estimating the ‘true’ effect (Borenstein, 2019). In meta-analysis we formally include sampling variance to remove this source of heterogeneity directly because it is known – hence why having sampling variances for an effect is so important (Gurevitch et al., 2018; Koricheva et al., 2013). Second, there are sociological factors impacting heterogeneity. These include attitudes towards publishing non-significant results. Low powered studies that do not find an ‘effect’ or that find an effect in the opposite direction often lead to what is called publication bias (Jennions et al., 2013; Nakagawa et al., 2021a; Rothstein et al., 2005). Third, there is real biology driving differences in effects. This is particularly true in comparative physiology where different species with varying habitats, life-histories, mating systems and reproductive modes are synthesised. As comparative physiologists, this is the stuff that gets us up every morning! Much variation could result from shared evolutionary history (i.e., phylogeny) or because of the different thermal environments species (or populations) occupy. We can model these biological factors to test predictions from hypotheses that we believe are key players impacting mean effect size and direction. This improves our understanding of the biological world around us and helps inform on future research directions. Finally, methodological factors are also big contributors to effect variability. In some cases, these methodological factors are of direct interest. For example, the methods for measuring an outcome variable can result in different effects, and it is important we know and understand this as it can determine how experiments are designed in the future. In other cases, methodological factors might be considered ‘nuisance heterogeneity’ (following from Cooper et al., 2019) – factors we know vary, but are rather unsurprising to us (Figure 1). These might include differences in temperature, pH, dosages, water potential etc. It might already be clear that these factors will have an impact on the magnitude of an effect, but they might distract from the core interest of a meta-analysis. In some cases, however, the latter factors might actually not be ‘nuisance’ variables but the main interest of a synthesis. Regardless, in many cases they can result in a more complicated interpretation of an effect measure than what is desired.

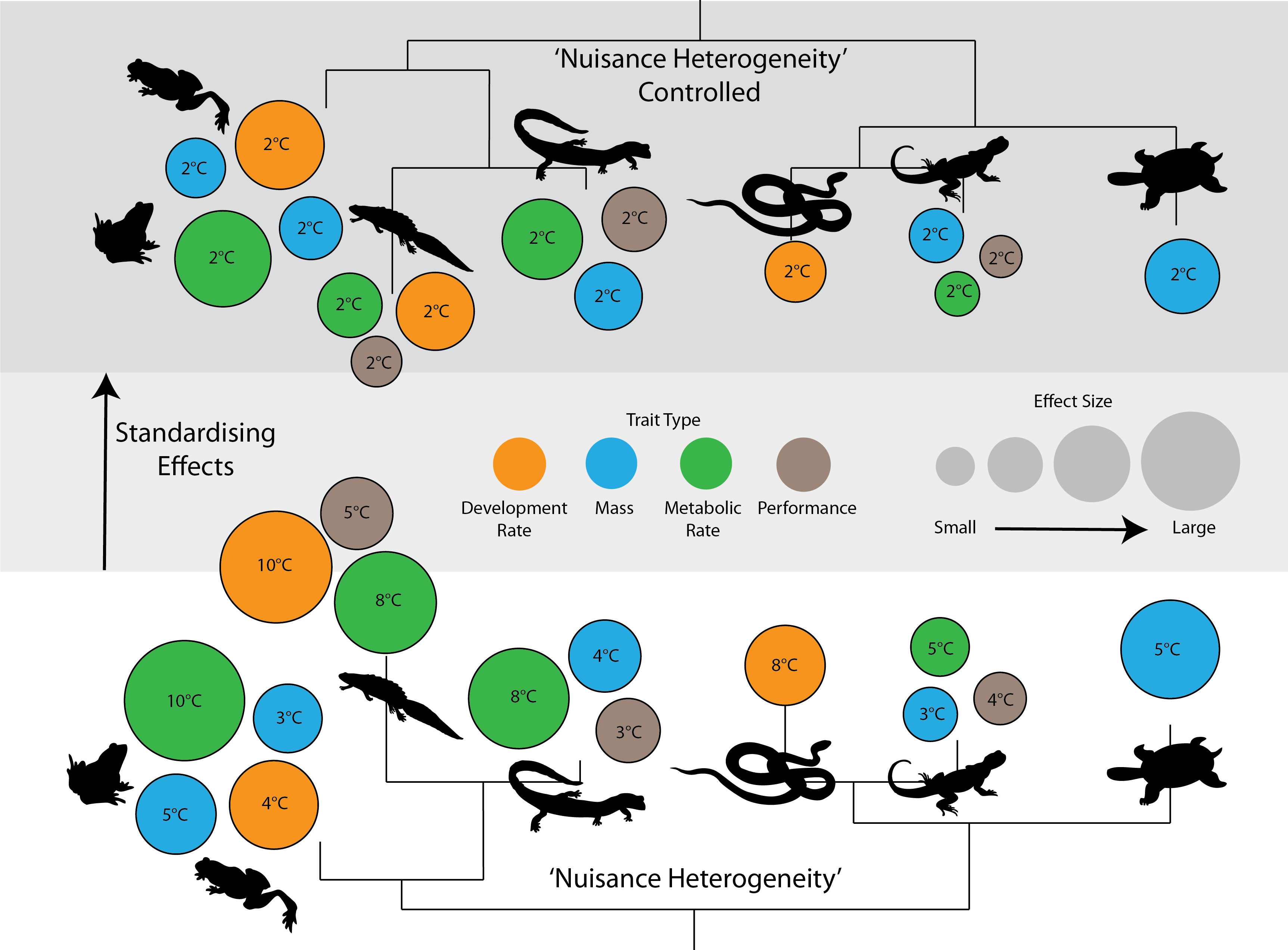


Figure 1: *Sources of effect size heterogeneity in meta-analyses*. Heterogeneity can come in various forms from effect measures varying as a result of shared evolutionary history (i.e., phylogeny), different trait types measured, and even sources of ‘nuisance heterogeneity’. Nuisance heterogeneity can be factors, such as temperature differences impacting effect size measures, or dosages. We can control for these types of heterogeniety by reformalising effect sizes and/or using multilevel metaregression as we describe in our review

All the various sources of heterogeneity (except sampling variance) contribute (to varying extents) to the ‘apples and oranges’ problem often touted as being an issues with meta-analyses (or quantitative syntheses more generally). Indeed, some of this criticism revolves around combining studies that vary in their quality (e.g., observational vs experimental). Combining standardised effects, while ignoring these sources of heterogeneity can result in the meaning of a meta-analysis not being particularly useful. However, many argue that this is just what we are interested in synthesizing (Borenstein, 2019; Cooper et al., 2019; Glass, 2015; Gurevitch et al., 2018). In the words of Gene Glass, the father of meta-analysis: “Of course it [meta-analysis] mixes apples and oranges; in the study of fruit nothing else is sensible; comparing apples and oranges is the only endeavor worthy of true scientists; comparing apples to apples is trivial” (p.g., 224, Glass, 2015). Glass (2015)‘s intention was to simply highlight that whether comparing apples and oranges is a good idea simply boils down to the specific question of interest, the population one wishes to make conclusions about and how heterogeneous effects actually are. If we are interested in gaining insight into ’fruit’ then synthesising apples and oranges is the only thing that makes sense. Instead, if we are interested in generalising only to apples then mixing fruit will not be ideal. In addition, if we want to understand why different ‘fruit’ vary in their effects, then we can only ask this question by including all fruit (Borenstein, 2019).

We tend to agree with Glass (2015), and his sentiment is of particular relevance to comparative physiology. For example, if we want to understand the impact of a pollutant on reproduction in aquatic organisms then including aquatic mammals and fish would provide answers to this general question, even if their reproductive biology and physiology are known to differ. By synthesizing diverse study outcomes, it allows us to not just report on overall effects (if the question is relevant to ask), but to understand critical features of the literature, study systems (e.g., different effects of pollutants on fish and mammal reproduction) and approaches that can explain the diversity of effect outcomes observed. In doing so, it provides a richer set of conclusions that can draw attention to important sources of variability that might have otherwise been overlooked. Having said that, it is still very important to understand the limitations and diversity of studies included within a meta-analysis, and how this can affect interpretations made – an important reason why transparency and reproducibility are such a prominent feature of meta-analysis (or ideally should be) (O’Dea et al., 2021). The validity of general conclusions to certain questions (e.g., what is the overall effect of a pollutant on reproduction?) will inevitably depend on the homogeneity of effects being synthesized. If the estimate of an overall effect coincides with low heterogeneity estimates then the effects are reasonably consistent (Borenstein, 2019). If, however, an overall effect is accompanied by large heterogeneity estimates, then we would shift focus and explore drivers of that heterogeneity. For example, given the very different physiology of aquatic mammals and fish, understanding an overall combined effect on reproduction may not be particularly relevant, and indeed may be misleading. Therefore, understanding overall effects in each of these groups separately (and whether they differ) will be more useful. It is important to consider these questions when conducting meta-analyses.

Of course, there are certain types of heterogeneity that transcend different questions or that are simply a ‘nuisance’ to interpreting effects. In other words, if we compare different types of ‘apples’ is it still fair to synthesize these studies to understand overall effects? This situation might be analogous to a meta-analysis in comparative physiology that attempts to understand the effects of diet on fish growth across different populations of a widely distributed (cosmopolitan) species. Studies might rear populations of fish under different temperatures, but in the end, they are still all the same fish species, governed by the same thermodynamic effects on growth. In this situation, we might still be interested in understanding overall effects, but we may question whether such a conclusion is sensible given the diversity of temperatures applied. It is not always clear how we should tackle these sources of heterogeneity. If temperature differences are of direct interest, then we should definitely model them. If, however, it is less interesting, than say the diet administered, then it may introduce an unnecessary complication to the interpretation of the effect(s) of interest. Here, we focus on meta-analytic tools to circumvent these issues and demonstrate that the ‘apples and oranges’ problem is not really a problem at all.

## Survey of Meta-Analytic Practices in Comparative Physiology

Before embarking on “meta-analytic enlightenment” it is worth first understanding how comparative physiologists are currently conducting meta-analyses / comparative analysis. As discussed above, different types of questions and experimental approaches will mix ‘apples and oranges’ to varying extents, and indeed, this heterogeneity (or basket of fruit) is often the main interest to comparative biologists. To better understand the different types of effect sizes being used, their susceptibility to known sources of methodological / environmental variability across studies, and how these sources are controlled we conducted a literature survey. To accomplish this goal, we searched for meta-analyses and comparative analyses that tackled questions using physiological data compiled from animal studies across multiple species. We were specifically interested in asking: 1) what type of effect size was used within the quantitative synthesis? 2) based on the methods provided, was the effect size likely susceptible to continuous variation in how treatments were applied (e.g., for contrast-based effects) or the conditions under which traits were measured (i.e., for correlations) across studies (i.e., ‘nuisance heterogeneity’)? 3) If so, was the variation accounted for in the effect size or during analysis? and finally, 4) how many analyses used ‘weighted’ meta-analytic models?

#### *Literature Search and Survey Data Collection*

All searches were conducted in Scopus between May 16 - 18, 2021. Briefly, we first did a series of pilot searches (detailed in Supplementary Materials) to refine our search query (See Foo et al., 2021 for an overview of this process). We searched for studies that included: ‘meta-analysis’, ‘meta regression’ , ‘comparative analysis’, ‘comprehensive analysis’ ‘global analysis’ or ‘macro physiology’ in their title and abstract, restricting our search to key comparative physiology journals (detailed in Table S1 and S2 - Supplementary Materials) within the last 6 years. We expected this to provide a contemporary overview of quantitative syntheses in comparative physiology while making the survey and screening feasible. We acknowledge that this may have missed some important, and influential, quantitative synthesis in the field of comparative physiology, but likely provided a random sample to inform on the questions we sought to address (see below).

We included quantitative syntheses, including meta-analyses (*sensu stricto*), that were focused on comparative physiological questions only. Narrative reviews and comments were not considered because they did not collate quantitative data from the literature. Studies needed to have collected physiological trait data from the literature including traits, such as hormones, thermal physiology (CTmax, CTmin, thermal preferences, thermal breadth, LT50, lethal thresholds), metabolic rates (routine, maximum, and minimum metabolic rate, aerobic scope), enzyme reaction rates and concentrations, oxidative stress and consequences (e.g., DNA damage, teleomere attrition), performance measurements (e.g., heart rates, sprint speed, swim speed, growth rates), body composition (e.g., fatty acid profiles) and immune traits (e.g., cell function, antibodies, ect.). Studies were also included if they measured the relationship between physiological traits and body size (e.g., metabolic scaling), reproductive allocation, or survival / mortality across taxa. Studies on behavior were only included if the synthesis was explicitly interested in the proximate physiological drivers of behavioral variation. We also included meta-analyses that quantified plasticity in physiological traits across taxa for the purpose of understanding adaptive plasticity in response to environmental variation. While many studies in comparative physiology collect data as part of a comparative analysis, we focus here only on studies collating data from existing literature because this is the major data source for quantitative synthesis. Quantitative syntheses interested in phenology, bio-mechanics, sexual size dimorphism (i.e., SSD, except if the study was interested in sexual differences in physiological traits), those focused on validating physiological methodology, or characterising population genetic structure, gene transcription and quantitative genetics were excluded. There were also meta-analyses that focused on colour evolution and contrast which varied in their goals but were most often not focused on physiological questions. As such, these studies were also excluded. We restricted our focus to animal studies because plant synthesis papers were often only focused on ecological (e.g., quantified changes in abundance and distribution of species) or agricultural questions (e.g., studies on soil microbial function, nitrogen and phosphorus capture, methane output and greenhouse gas emissions).

All title, abstract and full text screening was done by two people (DWAN and ML). Any conflicts or disagreements were resolved by discussion and a consensus on a paper was reached. Out of the 417 papers, 79 full texts were screened and we included a total of 62 for final data extraction. For full details on the search strings used, dates searched, the number of papers found, and our PRISMA flowchart (along with reasons for exclusion at the full text stage) we refer readers to the *Supplementary Materials*. To address our questions, we categorised effect sizes as falling into the following categories: Response Ratio (lnRR), Standardised Mean Differences (SMD; e.g., Hedges d, g or Cohen’s d), Correlation or Fisher’s Z -transformed Correlation, Regression Slopes (Slopes), Risk Ratios, or whether the paper analysed the raw data or summary statistics directly (i.e., means/proportions) (which we label as ‘raw’). For all other effects we labelled as ‘other’. To determine whether effect sizes were likely to vary based on some underlying continuous variable or were dependent on the context in which traits were measured, we assessed whether authors: 1) acknowledged this explicitly in the paper or 2) accounted for it in their models. If it was not clear, two of us (DWAN & PP) discussed the paper, and made a decision about whether the effect measure was likely to be impacted. For example, trait means, such as ectothermic metabolic rate (i.e., raw/mean/proportion-based effect size), will vary depending on the temperature being measured, the magnitude of differences between two treatments (e.g, contrast-based effect sizes - SMD or lnRR) might depend on the temperature or dosage used in treatments, and the strength of relationship between two variables (e.g., correlation between behaviour and metabolism) might depend on the temperature context under which metabolic rate or behaviour were measured. Given the high heterogeneity in effects used in comparative physiology, we expected ‘nuisance heterogeneity’ to be common to most effect sizes, however, it has not been formally quantified. We categorised the types of statistical models fit in each study as: 1) fixed-effect meta-analysis; 2) random effects meta-analysis (i.e., intercept only model, but has random study effects); 3) multi-level meta-analysis (i.e., intercept only model, but has multiple random effect levels such as study, species, observation-level random effects); 4) meta-regression models (i.e., a multilevel or random effect model with fixed effects - including random effects); 5) weighted regression models, including mixed effect models; and 6) linear regression models (including linear and generalised linear mixed effect models, and phylogenetic least squares models). Notably, only model types 1-5 are weighted models because they weight effect sizes based on the inverse of their sampling variance or sample size.

#### *Current Meta-analytic Practices in Comparative Physiology*

Quantitative syntheses in comparative physiology commonly analyse raw summary statistics (28.36% of studies), followed closely by Fishers transformed correlation coefficients (Zr) (23.88% of studies), standardised mean differences (SMD) and log response ratios (lnRR) (23.88% and 11.94%, respectively; Figure 2A). In many cases, studies make use of multiple effect size measures. Approximately 72.58% of effect sizes used within the study we deemed to be impacted by ‘nuisance heterogeneity’ in some form or another (Figure 2B), but of the studies where this was the case, only 53.33% of the studies clearly dealt with it (Figure 2B); usually through meta-regression modelling. Overall, 62.9% of studies used a weighted meta-analytic model, however, of the studies self described as ‘meta-analyses’, 17.39% (N = 8 of 46) did not actually weight their meta-analytic mean. Many studies used ‘Linear Regression’ (including linear and generalised linear mixed effects models) as the main type of model fit to the data (Figure 2C). Of the studies using weighted meta-analytic models 35% (N = 14 meta-analyses) did not report any measure of heterogeneity, which is unfortaunte given how important these measures are to interpreting effects.

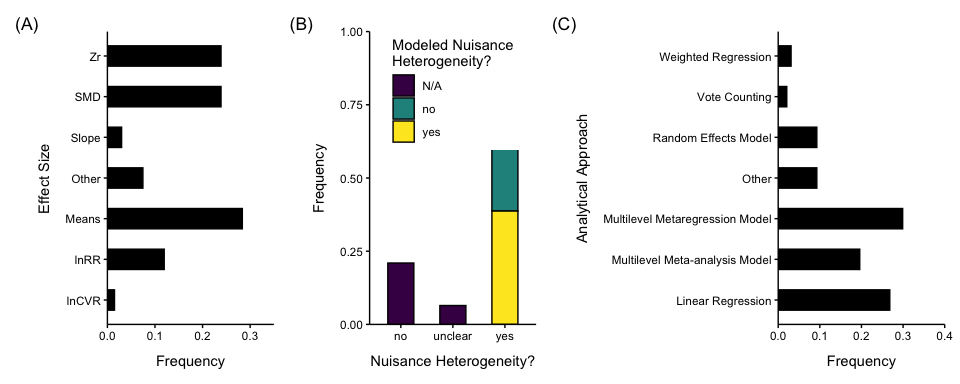


Figure 2: **Summary of meta-analytic practices in comparative physiology.** (A) Different types of effect measures commonly used in meta-analyses and their relative frequency of use. (B) how many meta-analyses where ‘nuisance heterogenity’ have the potential to impact effect sizes, and the proportion of studies explicitly dealing with nuisance heterogenity. (C) Frequency of analytical approaches used to analyse effect measures. Note that only ‘Weighted Regression’, ‘Random Effects Models’, ‘Multilevel Metaregression Models’ and ‘Multilevel Meta-analysis Models’ are weighted models.

## Opening Up New Opportunities in Comparative Physiology: Expanding the Breadth of Effect Sizes to Deal with “Nuisance Heterogeneity”

#### *Common Effect Sizes and Sampling Errors that are Useful in Comparative Physiology*

Meta-analysis requires comparative physiologists to make use of effect sizes that are suitable for answering their specific question. The key features of all effect sizes used in meta-analyses are that they are: 1) statistical parameters that have been placed on a common scale so that they can be compared across studies and 2) that they have some associated measure of precision (i.e., sampling variance)(Borenstein et al., 2009; Koricheva et al., 2013; Schmid et al., 2021). Effect sizes in comparative physiology are quite variable; they can be the mean difference between two groups (e.g., difference in mean between a control and treatment group - see Wu and Seebacher, 2020), the slope of a regression (e.g., between body size and metabolic rate), a correlation between two variables, or even the raw mean (or variance) itself (e.g., comparing CTmax or CTmin across species - see Sunday et al., 2014). In many cases, different effect sizes might be used depending on the question. For example, studies interested in the effect of temperature on some physiological trait that have measured the trait across studies experimentally manipulating temperature could analyse the mean ‘effect’ by creating contrast-based effect sizes or by analysing the mean in each temperature treatment. The choice as to which one to use will boil down to the types of data available from primary studies, the specific question of interest and whether studies are observational or experimental in nature. Considering our temperature example above, if we are interested in the magnitude of change between temperature treatments, we may want to use a contrast-based effect size, such as the log response ratio . If we are instead interested in how the mean of some physiological trait changes with temperature we may instead model the mean itself. In the latter scenario, so long as all the means are measured in the same units (e.g., ; or can be converted to the same units) then the mean may provide the desired answer to a given question. We provide details on different effect measures and associated sample sizes in Table 1.

Table 1: Common effect sizes used throughout meta-analyses in comparative physiology, their associated sampling variances and examples on when they might be used.

| **Effect Measure** | **Definition** | **Sampling Variance** | **Examples** |
| --- | --- | --- | --- |
| Mean |  |  | CTmin, CTmax, Metabolic Rate (MR) |
| Log Standard deviation, lnSD |  |  | Variability in CTmin, CTmax, MR |
| Log Response Ratio, lnRR |  |  | Ratio between BPA treatment (BPA-exposed group) and control (no BPA) |
| Standardised Mean Difference, SMDa |  |  | Difference in immune response between males and females or performance difference in the presence of stressor compared to absence of stressor |
| Zr (Fisher Trasnformaiton of   correlation Coefficient, r) |  |  | Relationship between sex hormones and immune responses or metabolic rate and behaviour |
| a; a | | | |

#### *Expanding the Scope of Effect Size Metrics in Comparative Physiology*

As can be seen in Table 1 none of the effect measures control for ‘nuisance heterogeneity’. For example, the magnitude of response between two treatment groups for lnRR or SMD might depend on just how different the temperatures, dosages, pH or water potential applied in each treatment. Even modelling just the mean often needs to control for various sources of ‘nuisance heterogeneity’. For example, standard metabolic rate (SMR) is temperature dependent. As such, one needs to standardise the temperature measurement to compare across studies (Uyeda et al., 2017; White et al., 2006) (if, of course, temperature is not of interest in of itself). While we could model this explicitly, this may not always be desirable. It can often be useful to standardise the effect to both ease its interpretation and simplify modelling. In Box 1 we provide readers with guidance on how comparative physiologists can apply corrections to the effect measure itself. We highlight two examples in the box focusing mainly on generality. Below, we show using a common, well understood effect measure of interest to comparative physiologists, , how we can formalise a traditional effect size, such as , to capture this ‘nuisance variation’.

#### *Comparing Changes in Mean Physiological Rates,*

A common experiment in comparative physiology is to manipulate the temperature organisms experience and measure some physiological rate (e.g., metabolic rate). Using the effect size measures in Table 1 would result in the effect sizes varying with the temperatures applied within and across across studies contributing to ‘comparability’ issues. For example, an effect size for a temperature manipulation of 10°C would be larger than an effect size for a temperature manipulation of 5°C (see Figure 1). As such, it is common to standardize effects by temperature. One common effect size measure already used in comparative physiology to compare physiological rates is (e.g., Rodgers et al., 2021; Havird et al., 2020; Seebacher et al., 2015), which describes the multiplicative change in physiological rates across a 10°C temperature change. Higher values indicate larger changes in physiological rates. Currently, however, we do not have associated sampling variances derived for making it challenging to make use of the full power of weighted meta-analytic models. Interestingly, can be seen as a variant of *lnRR*, meaning that we can derive a based effect size and sampling error using the well known mathematical properties of *lnRR*. This opens up the meta-analytic toolbox and improves upon our ability to account for well known sources of non-independence that are typical in meta-analysis (Lajeunesse, 2011; Noble et al., 2017).

Prior to showing how the relevant effect size can be calculated it is useful to understand its similarities to *lnRR*. The *lnRR* described by Hedges et al. (1999) and extended by Lajeunesse (2015) and can be calculated as follows (but see also Senior et al., 2020):

In equation (1), is the mean of group 1, often a control group, where as is the mean of group 2 (e.g., a treatment group). The mean of for group *i* can be any measurement type (e.g. a physiological rate, mass etc) so long as the measurement variable is ratio scale. Log transformation of this ratio makes the effect size normally distributed (most commonly assumed by meta-analytic models). Equation (2) is the analytical solution for *lnRR*’s sampling variance where, and is the sample variance in group 1 and 2, respectively and and is the sample size in group 1 and 2.

The equations for *lnRR* and its sampling variance allow us to easily extend this to based effect sizes. Recall that is described as follows:

Here, and are mean physiological rates and and are the temperatures that these rates are measured. Log transformation of equation (3) leads to the following log transformed :

Equation (4) is essentially a temperature corrected equivalent of lnRR when the numerator and denominator are measured at different temperatures. This allows one to compare the mean of two temperature treatments directly regardless of the temperatures that these groups have been measured. Here, we will refer to this as the log response ratio, . Notably, by taking the natural logarithm, there are two statistical advantages: 1) 0 becomes biologically meaningful as 0 means two rates are the exactly the same and ‘significantly different from 0’ means that two rates are significantly different to each other and 2) is more likely to satisfy the assumption of residual normality than (see Hedges et al. (1999)). The recognition of this equivalence means that we can easily calculate the sampling variance for equation (4) as follows:

#### *Formalising Effect Size Metrics that Compare Changes in Variability Across Treatments in the Presence of “Nuisance Heterogeneity”*

Nakagawa et al. (2015) recently proposed analogous effect size estimates to *lnRR* that allow for comparisons of changes in variance between two groups, the log variance ratio (*lnVR*) and the log coefficient of variation (*lnCVR*). Like *lnRR*, *lnVR* and *lnCVR* are ratios that describe the relative difference in trait variability between two groups. We refer readers to Nakagawa et al. (2015) for the equations describing *lnVR* and *lnCVR*, but these can easily be extended to their analogues (and associated sampling variances) as follows:

Equations (6) and (7) describe the change in physiological rate variance (eqn (6)) across a 10°C temperature change along with its sampling variance (eqn (7)). While this is a useful metric, as discussed by Nakagawa et al. (2015) there is often a strong mean-variance relationship that needs to be accounted for in analysing changes in variance. As such, we can calculate the coefficient of variation, which standardises changes in variance for changes in means as follows:

where is the coefficient of variation defined as . Whether one choose to use SD or CV-based effect size depends on questions at hand (See Nakagawa et al., 2015; Senior et al., 2020).

Our example using , a well known effect measure in comparative physiology, shows how common effect size’s can be re-formalised to account for ‘nuisance heterogeneity’, in this case temperature. Of course, in doing this we are making assumptions about the nature of temperature effects on an effect size. Nonetheless, we can apply similar approaches to other commonly used effect size metrics. We describe more generally how this can be done in Box 1, applying similar principles to standardise temperature differences for slopes and for standardised mean differences.

## Meta-Analytic Models to Control for and Investigate Heterogeneous Effects Within and Across Studies

It will not be possible to derive an effect size that completely controls for ‘nuisance’ heterogeneity at the effect size level all the time. This limitation is partly because treatment effects may not be simple linear or exponential functions of the mean response making the effect sizes we discuss above unsuitable (e.g., thermal performance curves; Noble et al., 2018). For example, dosages applied to treatments can follow quite complex non-linear relationships (e.g., quadratic relationships) in relation to some mean response. When this occurs we recommend comparative physiologists apply meta-regression approaches, controlling for ‘nuisance heterogeneity’ as a moderator. In many cases, this will be the easiest, or even preferred approach because exploring drivers of heterogeneity will be the main interest in most meta-analyses anyway. Nonetheless, there are some simple statistical approaches a meta-analyst can use to then make the overall effects more easily interpretable, allowing one to control for ‘nuisance heterogeneity’, while also testing other biological moderators of interest (Schielzeth, 2010). Before diving into meta-regression models, we will first overview the meta-analytic models commonly used.

#### *Multi-level Meta-Analysis when the Overall Effect is of Interest*

A common goal of meta-analysis is to obtain an overall meta-analytic mean effect size estimate and some form of uncertainty around that mean, like a 95% confidence interval (CI). Including a prediction interval (PI) can also be quite useful (see below for why) (Nakagawa et al., 2021b). Estimating an overall mean and 95% CI is generally done with a simple random effects or multilevel meta-analytic model that accounts for effect size sampling error (Nakagawa and Santos, 2012; Nakagawa et al., 2017). More often than not, meta-analyses in comparative physiology will have complex hierarchical structure. For example, multiple effect sizes may be taken from a single study on different traits measured on the same individuals, or come from many different species that share an evolutionary history (Cinar et al., 2021; Noble et al., 2017). As such, it is far more likely that comparative physiologists will need to apply a multi-level meta-analytic model to control for the various sources of non-independence, and whenever possible they should (Cinar et al., 2021; Nakagawa and Santos, 2012; Nakagawa et al., 2021c; Noble et al., 2017; Song et al., 2021). Multilevel models are also extremely useful in describing the various sources contributing to effect size variance (a feature we describe below). For the purpose of our discussion, we will assume the meta-analyst needs to use more advanced multilevel models. Having said that, it may be sufficient for a meta-analysis to make use of traditional random effects models. We refer readers to excellent reviews and books that overview the distinctions between common (fixed) effect, random effect and multilevel models and when one might want (or not want) to apply these models (Borenstein et al., 2009; Koricheva et al., 2013; Nakagawa and Santos, 2012; Nakagawa et al., 2017; Schmid et al., 2021).

Assume that we are interested in understanding the overall effect of, say, stress (i.e., corticosterone - CORT) on reproduction (e.g., see Podmokła et al., 2018). We might extract data from studies manipulating corticosterone in bird eggs. Of course, these studies are diverse, apply different dosages of corticosterone and are working across many species. If the studies are experimental in nature then a standardised effect size, such as or Hedges’ *d*, might be applicable. We then might fit an overall multilevel meta-analytic model as follows:

where is the *i*th effect size (*i* = 1 … , the number of effect sizes), is the overall meta-analytic mean, is the phylogenetic effect (a random effect) for species *k* applied to effect size *i* (i.e., species effects because of shared evolutionary history, Cinar et al., 2021). Phylogenetic effects are assumed to be normally distributed deviates sampled from a distribution with a mean of 0 and variance [i.e., *N*], where is a phylogenetic correlation matrix derived from a phylogenetic tree. is an additional species-specific effect for the *k*th species applied to effect size *i* (i.e., these are species-specific effects because of shared ecology and other factors, Cinar et al., 2021), is the study-specific effect for the *j*th study applied to effect size *i*, is the effect size-specific effect (or within study effect, or residuals) for the *i*th effect size, and is the sampling variance effect around the *i*th effect as a result of variation in precision among each effect sizes, , which is known as the sampling variance for effect. This model is a ‘weighted’ multilevel meta-analysis model because the estimation of the overall meta-analytic mean, , is weighted by the inverse sampling variance of each effect size. Importantly, if we have effect sizes that are derived from the *same* sample of animals (or plants) (e.g., may be because many traits are measured on the same sets of animals) then we need to remove this dependency from the calculation of each ; we can then model the known sampling covariance between effect sizes by modifying to include off-diagonals that account for this covariance (See Box 1 and Noble et al., 2017). The fact that we can obtain an overall estimate of sampling variance (i.e., , see below for how it is defined) is very important because it provides a way for us to understand just how much variation in effects are the result of sampling differences across studies versus other, potentially important sources of variation such as differences in the methods applied, species differences or even real biological effects that impact the effect size a study is likely to observe (Nakagawa and Santos, 2012; Nakagawa et al., 2017). This is formalised in what we call ‘heterogeneity analysis’ which we describe below. Importantly, this model provides us with the answer to our overall question: Across all the species and studies manipulating egg corticosterone, how much of an overall effect does corticosterone have on reproduction relative to control conditions?

Of course, the overall effect does not provide us a way to interpret its magnitude in the context of how much CORT has been applied in the treatments used in the sample of studies. The most we can say about it, is that the effect is small, moderate or large out of the sample of effect sizes extracted from this set of studies. In the next section, we show how to ‘correct’ the overall effect size so that its interpretation is more meaningful.

#### *Multilevel Meta-regression to Understand Variation and Improve Interpretation*

Heterogeneity among effects that result from ‘nuisance heterogeneity’ can be dealt with using multilevel meta-regression models. Multilevel meta-regression models include all the same random effects that we have already discussed, but they also include fixed effects (i.e., predictors or moderators) that attempt to explain changes in the mean effect size. In other words, we can modify our multilevel meta-analytic model and turn it into a multilevel meta-regression model as follows:

where is simply the sum of all effects for all moderators (fixed effects), . All other notation is the same as described above. Importantly, the variable could be any moderator collected from studies, including dosage, temperature, salinity, or pH. Some moderators will be at the level of the study (e.g., was the study experimental or observational in nature) but others could be taken at the effect size level (i.e., was the effect derived using a temperature of 23°C or 35°C). By including these moderators in a meta-regression model, the interpretation of the overall meta-analytic mean, , changes. For example, assume we collected the change in CORT dosage relative to a control group (i.e., in uL / g) for our meta-analysis on egg CORT effects on bird reproduction. If the untransformed dosage, , is included in the model then the overall meta-analytic mean estimate, , would be the mean effect of CORT relative to a control treatment, on reproduction when the dosage difference between the treatment and control group is 0 uL /g. Of course, this interpretation does not make much sense.

We can, however, apply some simple transformations to make the overall mean estimate more intuitive to understand with the added benefit that it can help improve the interpretation of non-linear parameters estimated by the model (Gelman and Hill; Schielzeth, 2010). One simple transformation would be to centre the variable , by subtracting each value from the mean to create a new input variable (i.e., , where the subscript *i* denotes each individual effect size dosage). The new variable, , is now centered on the mean and can replace in the meta-regression model. When the model is refit with the estimated can now be more intuitively interpreted as the overall mean effect of CORT on reproduction at an average dosage difference between treatment and control groups. Importantly, by centering, we retain the original units (uL / g) of the variable, which can be very useful to allow for a more general interpretation of the effect magnitude (e.g., when the dosage difference is 10 uL /g). Given it is very common to estimate and compare overall meta-analytic means (i.e., ) in meta-analyses this centering can be particularly useful in making these means more interpretable. Mean centering can even be extended to include other ‘nuisance’ variables that might creep into the dataset; the same centering can be done with a second variable, and the overall mean adjusted can be interpreted in the context of this new variable in addition to dosage. Of course, if one is not interested in the magnitude of the effect, we could also change our question and simply model mean reproductive output as a function of dosage (often what are referred to as ‘arm-based’ meta-analytic models). Here, one can model the mean for each group (i.e., control and treatment), and account for the known sampling variance associated with each group mean. Dosage, or other continuous variables like temperature, can then be modeled (accounting for sampling variance) to understand how mean reproductive output changes. It is important to recognise that all means need to be comparable and that these models are more complex to fit. While they can be equivalent to ‘contrast-based’ models, they require complex interactions to be estimated (Nakagawa et al., 2015).

Centering moderator variables, like temperature and dosage, not only provides a way to make the overall mean effect more interpretable in the face of treatment heterogeneity, but it also allows more complex relationships to be fit without compromising interpretations, such as when one includes non-linear parameters (e.g., polynomials) and interactions with other variables. In other words, linear / main effects can still be interpreted normally in the presence of non-linear / interactions fit in the model (Gelman and Hill; Schielzeth, 2010). We provide examples in the online supplementary material (with R code) to show readers how the models we describe above can be fit, and how, meta-analytic means can be adjusted for treatment heterogeneity.

#### *Heterogeneity Analysis: How Much Effect Size Variation Actually Exists?*

As we have already emphasised, quantifying effect size heterogeneity (i.e., effect size variance) provides the impetus for exploring hypotheses that are thought to explain why effects might vary within and across studies, while also providing context for interpreting the generality of effects. There are a number of measures commonly used to quantify effect size heterogeneity (Borenstein, 2019), however, in ecological and evolutionary meta-analysis, heterogeneity is often calculated and reported as as follows (if using a multilevel meta-analytic model as we describe above):

where is the total effect size variance and is the total sampling error variance calculated as:

where *k* is the number of studies and the weights, , can be calculated using the inverse of the sampling variance () for each effect size, *i*.

Quantifying heterogeneity using provides a formal way to assess the relative amount of variation that is the result of ‘true’ biological or methodological differences as opposed to simply chance (i.e., sampling variance) (Higgins and Thompson, 2002; Nakagawa and Santos, 2012). In other words, can be thought of as the percentage of variance between effect sizes after removing the effects of sampling error (Higgins and Thompson, 2002). We can only calculate this metric if we have an estimate of the sampling variance for a given effect measure because only then can we quantify the proportion of variation that is the result of sampling variability. Senior et al. (2016) have shown that total heterogeneity is often extremely high in ecological and evolutionary meta-analysis (~91%), suggesting that effect measures vary widely across and within studies. Some of the variation is clearly the result of working with different species or strains, different traits and different outcome measures, however, some of this variation also arises because of methodological differences across studies that might be related to the dosages or temperatures used for treatments.

Formally quantifying and presenting heterogeneity estimates from meta-analytic models can provide a way to understand the major drivers of effect size variation by allowing researchers to produce various estimates which can be compared within and across studies. For example, we could fit the multilevel model described above to understand what proportion of variation is the result of, say, ‘study’ effects following Nakagawa and Santos (2012): where is the study specific variance.

Many readers will notice the similarities between and (Nakagawa and Schielzeth, 2010; Nakagawa and Schielzeth, 2013). If we want to formally assess how much effect size variation is the result of ‘nuisance’ variation, such as temperature and dosage differences applied across studies, then we could fit our multilevel meta-regression model, including dosage or temperature (linear, quadratic or even more complex fits), and estimate how much variation is explained by these fixed effects following Nakagawa and Schielzeth (2013).

Note that this formula does not include as sampling error variance is assumed to be known in meta-analysis as explained above. When including continuous moderators, such as dosage, temperature, salinity, and even pH, provides a formal means to assess just how much variation in effects is the result of ‘nuisance’ heterogeneity at the effect size level. Other biological moderators of interest could also be included as fixed effects and calculated for each independently or all together. Importantly, moderators could be those that explain variation at the effect size level (i.e., dosage, temperature) or they could explain variation at the study or species level (e.g., reproductive mode, endothermy etc). , or how we like to think of it in this context as , maybe a useful measure to help readers understand just how much effect variation can result from the specific treatment application. If the heterogeneity is high as a result of, say, dosages differences, then it is clear that the choice of dosage will have a critical impact on the effect of interest. Comparative physiologists interested in implementing these calculations can do so using our packages orchaRd (Nakagawa et al., 2021b) or metaAidR (<https://github.com/daniel1noble/metaAidR>).

Heterogeneity described using measures can be useful for understanding the relative contributions of different factors to effect size variation, however, prediction intervals might be more appropriate in many cases (Borenstein, 2019; Nakagawa et al., 2021b). Unlike , prediction intervals (PI’s) are probably more meaningful in the context of meta-analysis because they explicitly incorporate measures of dispersion around a mean effect size providing information about the likely effect size one can expect if we were to randomly sample a new population. For example, assume that we were interested in the overall impact of salinity stress in freshwater fish. To tackle this question, we collected experimental studies measuring fish swim performance under high salinity treatments (~15-20ppm Na Cl) compared to freshwater (~0ppm Na Cl) controls; using SMD to quantify the magnitude of effect salinity had on swim performance. We conduct a meta-analysis and estimate an overall SMD of 0.50 with a prediction interval of 0.05 to 0.85. The PI indicates that effects can vary widely from as low as an SMD of 0.05 to as high as an SMD 0.95, depending on the population. In addition, if we were to repeat a study we would expect a new effect to fall within the range of 0.05 and 0.85, 95% of the time (Borenstein, 2019; Nakagawa et al., 2021b). This is a very intuitive interpretation of how variable effects are, yet PIs are often not reported in meta-analyses (Nakagawa et al., 2021b), and we encourage more meta-analysts to report these.

## Conclusions

Comparative physiologists are interested in meta-analysing effects across a diversity of species, experimental approaches / designs and environments. While an overall meta-analytic mean may be of interest, more often than not, understanding the drivers of effect size heterogeneity will be the primary interest. However, obvious drivers of effect size heterogeneity, such as temperature and dosage differences across studies, may not really be of interest. Comparative physiologists are likely to want an effect size or analytic approach that allows them to remove these ‘nuisances’. Here, we provide a set of tools to deal with ‘nuisance heterogeneity’ by ‘reinventing’ standard effect sizes or using meta-regression models. Of course, ‘nuisance heterogeneity’ might in of itself be the main interest, in which case standard meta-regression approaches will work sufficiently. However, in many cases, comparative physiologists might simply want to focus on bigger, more interesting biological drivers of effect size direction and magnitude. Estimating complex meta-regression models with a multitude of fixed effects might result, however, in more challenging model to interpret As such, a combined approach, where some of the known heterogeneity is dealt with by deriving a variant effect size and associated sampling variance, and other biologically interesting moderators are fit in a meta-regression model, might be desired. Indeed, this may even be a necessity to simplify the model. We hope the set of tools we describe provide clarification around different sources of heterogeneity; providing new ways to help comparative physiologists communicate effect measures more richly to guide our understanding and decisions around pressing global challenges.

## Awknowledgements

We would like to thank the Editorial board (Craig Franklin, Michaela Handel and Charlotte Rutledge) for inviting us to contribute to this special issue along with feedback on previous drafts. We would also like to than Megan Head for comments on a previous draft. DWAN and SN were supported by ARC Discovery Projets (DP210101152 to DWAN and DPXX to SN).

## References

**Arnqvist, G. and Wooster, D.** (1995). Meta-analysis: Synthesizing research findings in ecology and evolution. *Trends in Ecology and Evolution* **10**, 236–240.

**Borenstein, M.** (2019). Heterogenity in meta-analysis. In (ed. Cooper, H.), Hedges, L. V.), and Valentine, J. C.), pp. 454–466. New York: Russell Sage Foundation.

**Borenstein, M., Hedges, L., Higgens, J. and Rothstein, H.** (2009). Introduction to meta-analysis. *John Wily & Sons, Ltd, West Sussex, UK*.

**Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. and Munafò̀, M. R.** (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365–376.

**Carpenter, C.** (2020). Meta-analyzing apples and oranges: How to make applesauce instead of fruit salad. *Human Communication Research* **46**, 322–333.

**Cinar, O., Nakagawa, S. and Viechtbauer, W.** (2021). Phylogenetic multilevel meta-analysis: A simulation study on the importance1of modeling the phylogeny. *EcoEvoRxiv, https://ecoevorxiv.org/su4zv/*.

**Cooper, H., Hedges, L. V. and Valentine, J. C.** (2009). The handbook of research synthesis and meta-analysis. New york. Russell sage foundation.

**Cooper, H., Hedges, L. V. and Valentine, J. C.** (2019). Potentials and limitations of research synthesis. In (ed. Cooper, H.), Hedges, L. V.), and Valentine, J. C.), pp. 517–525. New York: Russell Sage Foundation.

**Foo, Y. Z., O’Dea, R. E., Koricheva, J., Nakagawa, S. and 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* **00**, 1–16.

**Forstmeier, W., Wagenmakers, E. and Parker, T.** (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews* 1941–1968.

**Gallo, P.** (1978). Meta-analysis: A mixed meta-phor? *American Psychologist* **33**, 515–517.

**Gelman, A. and Hill, J.** Data analysis using regression and multilevel/hierachical models. *Cambridge University Press, Cambridge, UK*.

**Glass, G.** (2015). Meta-analysis at middle age: A personal history. *Research Synthesis Methods* **6**, 221–231.

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

**Gurevitch, J., Koricheva, J., Nakagawa, S. and Stewart, G.** (2018). Meta-analysis and the science of research synthesis. *Nature* **555**, 176–182.

**Havird, J. C., Neuwald, J. L., Shah, A. A., Mauro, A., Marshall, C. A. and Ghalambor, C. K.** (2020). Distinguishing between active plasticity due to thermal acclimation and passive plasticity due to q10 effects: Why methodology matters. *Functional Ecology* **0**, 1–14.

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

**Higgins, J. P. T. and Thompson, S. G.** (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539–1558.

**Jennions, M. and Møller, A.** (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology* **14**, 438–445.

**Jennions, M., Lortie, C., Rosenberg, M. and Rothstein, H.** (2013). Publication and related biases. *In: Handbook of Meta-Analysis in Ecology and Evolution (eds J. Koricheva, J. Gurevitch & K. Mengersen). Princeton University Press, Princeton and Oxford.* pp 207–236.

**Koricheva, J., Gurevitch, J. and Mengersen, K.** (2013). Handbook of meta-analysis in ecology and evolution. *Princeton University Press, Princeton, New Jersey.*

**Lajeunesse, M. J.** (2011). On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology* **92**, 2049–2055.

**Lajeunesse, M. J.** (2015). Bias and correction for the log response ratio in ecological meta-analysis. *Ecology* **96**, 2056–2063.

**Morrissey, M.** (2016). Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *Journal of Evolutionary Biology* **29**, 1862–1904.

**Nakagawa, S. and Santos, E. S.** (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology* **26**, 1253–1274.

**Nakagawa, S. and Schielzeth, H.** (2010). Repeatbility for gaussian and non-gaussian data: A practical guide for biologists. *Biological Reviews* **85**, 935–956.

**Nakagawa, S. and Schielzeth, H.** (2013). A general and simple method for obtaining r2 from generalised linear mixed effects models. *Methods in Ecology and Evolution* **4**, 133–142.

**Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M. and Senior, A. M.** (2015). Meta-analysis of variation: Ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution* **6**, 143–152.

**Nakagawa, S., Noble, D. W. A., Senior, A. M. and Lagisz, M.** (2017). Meta-evaluation of meta-analysis: Ten appraisal questions for biologists. *BMC Biology* 15–18, DOI10.1186/s12915–017–0357–7.

**Nakagawa, S., Lagisz, M., Jennions, M. D., Koricheva, J., Daniel W.A. Noble, T. H. P., Sánchez-Tójar, A., Yang, Y. and O’Dea, R. E.** (2021a). Methods for testing publication bias in ecological and evolutionary meta-analyses. *EcoEvoRxiv* **10.32942/osf.io/k7pmz**,

**Nakagawa, S., Lagisz, M., O’Dea, R. E., Rutkowska, J., Yang, Y., Noble, D. W. and Senior, A. M.** (2021b). The orchard plot: Cultivating forest plots for use in ecology, evolution and beyond. *Research Synthesis Methods* **12**, 4–12.

**Nakagawa, S., Senior, A. M., Viechtbauer, W. and Noble, D. W.** (2021c). An assessment of statistical methods for non-independent data in ecological meta-analyses. *EcoEvoRxiv, 10.32942/osf.io/w96nv*.

**Noble, D. W., Lagisz, M., O’Dea, R. E. and Nakagawa, S.** (2017). Non‐independence and sensitivity analyses in ecological and evolutionary meta‐analyses. *Molecular Ecology* **26**, 2410–2425.

**Noble, D. W. A., Stenhouse, V. and Schwanz, L. E.** (2018). Developmental temperatures and phenotypic plasticity in reptiles: A systematic review and meta-analysis. *Biological Reviews* **93**, 72–79.

**O’Dea, R. E., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W. A., Parker, T. H., Gurevitch, J., Page, M. J., Stewart, G., Moher, D., et al.** (2021). Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: A prisma extension. *Biological Reviews* **doi: 10.1111/brv.12721**,

**Podmokła, E., Drobniak, S. and Rutkowska, J.** (2018). Chicken or egg? Outcomes of experimental manipulations of maternally transmitted hormones depend on administration method – a meta-analysis. *Biological Reviews* **93**, 1499–1517.

**Rodgers, E. M., Franklin, C. E. and Noble, D. W. A.** (2021). Diving in hot water: A meta-analytic review of how diving vertebrate ectotherms will fare in a warmer world. *Journal of Experimental Biology* **224**, 1–12.

**Rothstein, H., Sutton, A. and Borenstein, M.** (2005). Publication bias in meta-analysis: Prevention, assessment and adjustments. *Wiley, Chichester*.

**Schielzeth, H.** (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution* **1**, 103–113.

**Schmid, C. H., Stijnen, T. and White, I. R. eds.** (2021). *Handbook of meta-analysis*. Chapman & Hall CRC Press, Taylor & Francis Group.

**Seebacher, F., White, C. R. and Franklin, C. E.** (2015). Physiological plasticity increases resilience of ectothermic animals to climate change. *Nature Climate Change* **5**, 61–66.

**Senior, A., Grueber, C., Kamiya, T., Lagisz, M., O’dwyer, K., Santos, E. and Nakagawa, S.** (2016). Heterogeneity in ecological and evolutionary meta‐analyses: Its magnitude and implications. *Ecology* **97**, 3293–3299.

**Senior, A. M., Viechtbauer, W. and Nakagawa, S.** (2020). Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Research Synthesis Methods* **11**, 553–567.

**Song, C., Peacor, S. D., Osenberg, C. W. and Bence, J. R.** (2021). An assessment of statistical methods for nonindependent data in ecological meta-analyses. *Ecology* **e03184**,

**Stewart, G.** (2010). Meta-analysis in applied ecology. *Biology Letters* **6**, 78–81.

**Sunday, J. M., Bates, A. E., Kearney, M. R., Colwell, R. K., Dulvy, N. K., Longino, J. T. and Huey, R. B.** (2014). Thermal-safety margins and the necessity of thermoregulatory behavior across latitude and elevation. *Proceedings of the National Academy of Sciences (USA)* **111**, 5610–5615.

**Uyeda, J. C., Pennell, M. W., Miller, E. T., Maia, R. and McClain, C. R.** (2017). The evolution of energetic scaling across the vertebrate tree of life. *The American Naturalist* **190**, 185–199.

**White, C. R., Phillips, N. F. and Seymour., R. S.** (2006). The scaling and temperature dependence of vertebrate metabolism. *Biology Letters* **2**, 125–127.

**Wu, N. C. and Seebacher, F.** (2020). Effect of the plastic pollutant bisphenol a on the biology of aquatic organisms: A meta‐analysis. *Global Change Biology* **26**, 3821–3833.