REVIEW

Meta-analytic approaches and effect sizes to account for ‘nuisance heterogeneity’ in comparative physiology

Daniel W. A. Noble[1,*], Patrice Pottier[2], Malgorzata Lagisz[2], Samantha Burke[2], Szymon M. Drobniak[2], Rose E. O’Dea[2] and Shinichi Nakagawa[2]

ABSTRACT

Meta-analysis is a powerful tool used to generate quantitatively informed answers to pressing global challenges. By distilling data from broad sets of research designs and study systems into standardised effect sizes, meta-analyses provide physiologists with opportunities to estimate overall effect sizes and understand the drivers of effect variability. Despite this ambition, research designs in the field of comparative physiology can appear, at the outset, as being vastly different to each other because of ‘nuisance heterogeneity’ (e.g. different temperatures or treatment dosages used across studies). Methodological differences across studies have led many to believe that meta-analysis is an exercise in comparing ‘apples with oranges’. Here, we dispel this myth by showing how standardised effect sizes can be used in conjunction with multilevel meta-regression models to both account for the factors driving differences across studies and make them more comparable. We assess the prevalence of nuisance heterogeneity in the comparative physiology literature – showing it is common and often not accounted for in analyses. We then formalise effect size measures (e.g. the temperature coefficient, Q_{10}) that provide comparative physiologists with a means to remove 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 applied to a variety of different contexts to derive new effect sizes and sampling variances, opening up new possibilities for quantitative synthesis. 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 while making results from large-scale data sets more accessible, comparable and widely interpretable.

KEY WORDS: Multilevel meta-analysis, Quantitative synthesis, ‘Apples and oranges’, Sampling error, log Response ratio, Standardised mean difference

Introduction

Meta-analysis has emerged as the ‘gold standard’ across various disciplines for quantitative synthesis and is becoming increasingly prominent in the field of comparative physiology (Glass, 2015; Gurevitch et al., 2018; Nakagawa et al., 2017a,b). Meta-analyses answer research questions by synthesising research results and identifying sources of variation across studies (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., 2017a,b). Ideally, meta-analyses are part of a systematic review. As such, methods are carefully reported upon to ensure the review and data collation are transparent and reproducible [see the Preferred Reporting Items for Systematic Reviews and Meta-Analyses in Ecology and Evolutionary Biology (PRISMA-EcoEvo) checklist; O’Dea et al., 2021].

A meta-analysis can have three goals: (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) to attempt to identify research 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., 2017a,b). Meta-analysing independent studies provides greater statistical power and precision than what any individual study on its own would be able to provide – 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), we 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). Meta-analyses have already 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 bisphenol A (BPA), impact aquatic organisms (e.g. Wu and Seebacher, 2020).

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 are not comparable (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 (Arnqvist and Wooster, 1995; Stewart, 2010). For example, experimental designs in the field of comparative physiology can vary greatly in the temperatures or dosages of chemicals that they use 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 across studies. However, to many, having highly heterogeneous effects ‘…is the spice of life’ (p. 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).

[1]Division of Ecology and Evolution, Research School of Biology, The Australian National University, Canberra, ACT 2600, Australia. [2]Ecology & Evolution Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia.

*Author for correspondence (daniel.noble@anu.edu.au)

D.W.A.N., 0000-0001-9460-8743; P.P., 0000-0003-2106-6597; M.L., 0000-0002-3993-6127; S.M.D., 0000-0001-8101-6247; R.E.O., 0000-0001-8177-5075; S.N., 0000-0002-7765-5182

The goal of our Review is to briefly overview the meta-analytic process and argue for the value of heterogeneity in answering fundamental research problems and guiding research directions. We present results from 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. Then, we show how nuisance heterogeneity in comparative physiology can be comfortably dealt with by re-formalising many existing effect size measures and/or by using multilevel meta-regression models. We formalise alternative effect sizes and their associated sampling variance to provide comparative physiologists with opportunities to explicitly incorporate nuisance heterogeneity at the effect size level to ease their interpretation. Finally, we describe how more complex treatment differences, such as non-linear dosage differences, can be accommodated using multilevel meta-regression models. We hope that expanding the meta-analytic toolkit will provide new opportunities for comparative physiologists to address how organisms will cope with rapidly changing environments and anthropogenic stressors 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 of effects within and across studies (Borenstein, 2019; Cooper et al., 2019; Gurevitch et al., 2018; Lajeunesse, 2010; Nakagawa et al., 2017a,b). The concept of heterogeneity is vitally important because it tells us how general our findings are likely to be and how much of the variance we see is the result of real biology or methodological differences (after accounting for sampling variance) (Borenstein, 2019; Lajeunesse, 2010; Nakagawa et al., 2017a,b).

There are many phenomena driving effect heterogeneity. First, there is inherent uncertainty in estimating the 'true' population effect (e.g. a correlation coefficient) from smaller samples (Borenstein, 2019). A formal meta-analysis weights sample estimates by their precision, and removes this source of heterogeneity (Gurevitch et al., 2018; Koricheva et al., 2013). Removing sampling variance is possible because we know, mathematically, how to calculate it for many effect measures. While unweighted analyses are common, weighting studies makes overall estimates more precise, and less susceptible to publication bias (see below), even if the overall mean itself is unbiased (Morrissey, 2016). Second, there are sociological factors impacting heterogeneity. These include the ease with which novel and significant results are published relative to non-significant results, manifesting as publication bias (Jennions et al., 2013; Nakagawa et al., 2021a; Rothstein et al., 2005). Third, real biological processes can drive effect heterogeneity. This is particularly true in comparative physiology, where we synthesise data from different species with varying life histories, habitats, mating systems and reproductive modes. As comparative physiologists, this is the stuff that gets us up every morning! We can model these biological factors by including relevant predictors (moderator variables) and/or random effects (to account for non-independence of effects). By doing so, we can test predictions from hypotheses about the key players impacting mean effect size and direction. This improves our understanding of the biological world around us and helps inform 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 to determine how experiments can be 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 (Fig. 1). These might include differences in temperature, pH, dosages, water potential, etc. Regardless of the relative importance of such variables compared with core questions in a meta-analysis, nuisance variables may complicate the interpretation of effect sizes used if not accounted for properly.

Besides sampling variance, all the sources of heterogeneity discussed above contribute (to varying extents) to the 'apples and oranges' problem often touted as compromising the reliability of meta-analyses. Combining standardised effects while ignoring these sources of heterogeneity can indeed weaken the reliability of a meta-analysis. However, many argue that heterogeneity is just what we are interested in synthesising (Borenstein, 2019; Cooper et al., 2019; Glass, 2015; Gurevitch et al., 2018; Lajeunesse, 2010). In the words of Gene Glass, the founder 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. 224, Glass, 2015). 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 generalising only to apples, then mixing fruit will not be ideal. However, if we want to understand why different 'fruit' vary in their properties, then we need to consider apples and oranges together (Borenstein, 2019).

We tend to agree with Glass (2015), and his sentiment is particularly relevant to comparative physiology. For example, if we want to understand the impact of a pollutant on reproduction in aquatic organisms, including aquatic mammals and fish would provide answers to this general question, even if their reproductive biology and physiology are different. Synthesising diverse study outcomes allows us to also understand critical features of the literature, study systems (e.g. different effects of pollutants on fish and mammal reproduction) and approaches (e.g. short-term versus long-term studies) that can explain the diversity of effect outcomes observed. Doing so provides a rich 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 in a meta-analysis, and how this can affect interpretations made – an important reason why transparency and reproducibility are such prominent features 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 synthesised. If the estimate of an overall effect coincides with low heterogeneity then the effects are reasonably consistent (Borenstein, 2019). If, however, an overall effect is accompanied by large heterogeneity, 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 may even be misleading. It may, therefore, be more useful to understand overall effects in each of these groups separately, and assess the extent to which they differ.

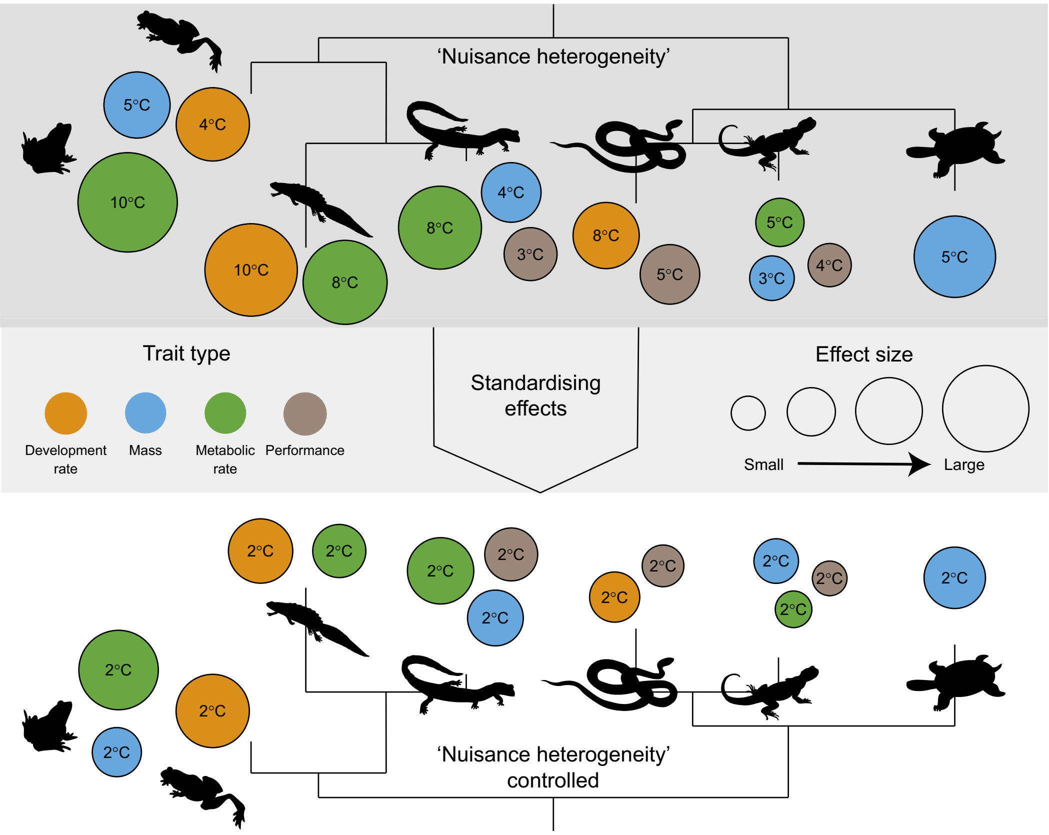

Fig. 1. Sources of effect size heterogeneity in meta-analyses. Heterogeneity can come in various forms, from effect measures varying as a result of shared evolutionary history (i.e. phylogeny), the different trait types measured, and sources of 'nuisance heterogeneity'. In this example, nuisance heterogeneity is generated by differences in the experimental temperatures at which the animals were tested (grey shaded area on top). However, nuisance heterogeneity can come in various other forms such as differences in dosage. We can control for these types of heterogeneity by reformalising effect sizes and/or using multilevel metaregression (white area on the bottom).

Nuisance heterogeneity complicates meta-analytic interpretations

Nuisance heterogeneity can get in the way of understanding real biological patterns that interest comparative physiologists. Take, for example, a meta-analysis that attempts to assess the effects of diet on fish growth across different populations of a widely distributed (cosmopolitan) species. Studies might rear populations of the same fish species under different temperatures, but in the end, they are still all governed by the same thermodynamic effects on growth. In this situation, we are still interested in understanding overall diet effects, but we may question whether such a conclusion is sensible given the diversity of temperatures applied. If temperature differences are of direct interest, then this variable can be formally incorporated in statistical models. If, however, it is not the primary interest, then it may introduce an unnecessary complication to the interpretation of the effect of diet on fish growth. Such heterogeneity is an important contributing factor to the 'apples and oranges' problems described above. Nonetheless, there is limited discussion about the ways in which it can be dealt with in the literature.

Here, we focus on meta-analytic solutions to circumvent nuisance heterogeneity, showing how it can easily be overcome using a number of meta-analytic tools. Prior to overviewing the solutions, we conducted a literature survey to better understand the different types of effect sizes being used in comparative physiology, the susceptibility of studies to nuisance heterogeneity, and how these sources are currently (if at all) being controlled for in a meta-analysis. More specifically, we asked: (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 nuisance heterogeneity?; (3) if so, was such variation accounted for in the effect size or during analysis?; and finally, (4) how many analyses used 'weighted' meta-analytic models?

We searched the Scopus database between 16 and 18 May 2021, with a follow-up search on 1 September 2021 (see the supplementary information at https://daniel1noble.github.io/nuisance_heterogeneity/). After a series of pilot searches to refine our search string (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 or abstract. We restricted our search to key comparative physiology journals known to the authors that would likely contain literature-based quantitative syntheses within the last 6 years (detailed in the supplemental tables at https://daniel1noble.github.io/nuisance_heterogeneity/). 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 syntheses in the field of comparative physiology, but likely provided an informative random sample.

All title, abstract and full text screening was done by two people (D.W.A.N. and M.L.). Any disagreements were resolved by discussion to reach a consensus. Out of the 426 papers originally identified, 80 full texts were screened for eligibility, and we included a total of 63 for final data extraction. For full details on the search strings used, search dates, 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 information (https://daniel1noble.github.io/nuisance_heterogeneity/).

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 (Z_r), regression slopes (slopes), risk ratios, or whether the paper analysed the raw data or summary statistics directly (i.e. means/proportions) (which we labelled as 'raw'). All other effects were labelled as 'other'. To determine whether effect sizes were impacted by nuisance heterogeneity, 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 (D.W.A.N. and P.P.) 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) will depend on the pH or dosage used in treatments, and the strength of the 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.

We also categorised the types of statistical models fitted 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) multilevel meta-analysis (i.e. intercept-only model, but has multiple random effect levels such as study-, species- and observation-level random effects); (4) meta-regression models (i.e. a multilevel or random-effects model with fixed effects – including random effects); (5) weighted regression models (including mixed effects models); and (6) linear regression models (including linear and generalised linear mixed-effect models, and phylogenetic least squares models) (see Nakagawa and Santos, 2012 for discussion of 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.

We summarise the results of our literature survey below while discussing solutions for dealing with nuisance heterogeneity.

Opening 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

The key features of all effect sizes used in meta-analyses are that: (1) they are statistical parameters that have been placed on a common scale so that they can be compared across studies and (2) 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 diverse; they can be the mean difference between two groups (e.g. difference in mean between a control and treatment group; e.g. Wu and Seebacher, 2020), the slope of a regression (e.g. between body size and metabolic rate; Uyeda et al., 2017), a correlation between two variables (e.g. between hormone levels and behaviour; Holtmann et al., 2016), or even the raw mean (or variance) itself (e.g. comparing CT_{max} or CT_{min} across species; e.g. Sunday et al., 2014).

In many cases, different effect sizes might be used depending on the question. For example, to investigate the effect of temperature on a physiological trait, one can compile studies experimentally manipulating temperature and analyse the mean 'effect' by creating contrast-based effect sizes, or alternatively, just model the trait in each temperature treatment. The choice as to which one to use boils 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 were 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, lnRR. Instead, if we were interested in how the mean of some physiological trait changes with temperature, we may model the mean itself. In the latter scenario, if all the means are measured in the same units (e.g. °C) or can be converted to the same units, then the mean may provide the desired answer to the original question.

The results of our literature survey show that quantitative syntheses in comparative physiology commonly analyse raw summary statistics (28.36% of studies), followed closely by Fisher's transformed correlation coefficients (Z_r) (23.88% of studies), standardised mean differences (SMD) (23.88% of studies) and log response ratios (lnRR) (11.94%, of studies) (Fig. 2A). We provide details on different effect measures and associated sample sizes in Table 1. In many cases, studies make use of multiple effect size measures. Approximately 72.58% of effect sizes used within the study were deemed to be impacted by nuisance heterogeneity in some form or another (Fig. 2B), but of the studies where this was the case, only 53.33% of the studies clearly dealt with it (Fig. 2B), often through meta-regression modelling.

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 – a point reinforced by our survey. For example, the magnitude of lnRR or SMD can depend on differences in the temperatures, dosages or pH applied to each treatment. The same applies to modelling the means. 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 itself). Although we could model nuisance heterogeneity 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 the Appendix, we provide readers with guidance on how comparative physiologists can apply corrections to the effect measure itself. We highlight two examples in the Appendix that can be easily generalized to alternative effect size measures and show how they are distinct from traditional effect measures (Fig. 3). Below, we focus on one familiar example to elaborate on how a traditional effect size, such as lnRR, can be reformalised to capture nuisance variation. We use a common, well-understood effect measure of interest to comparative physiologists – the temperature coefficient (Q_{10}).

Comparing changes in mean physiological rates, Q_{10}

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 from Table 1 would result in the effect sizes varying with the temperatures applied

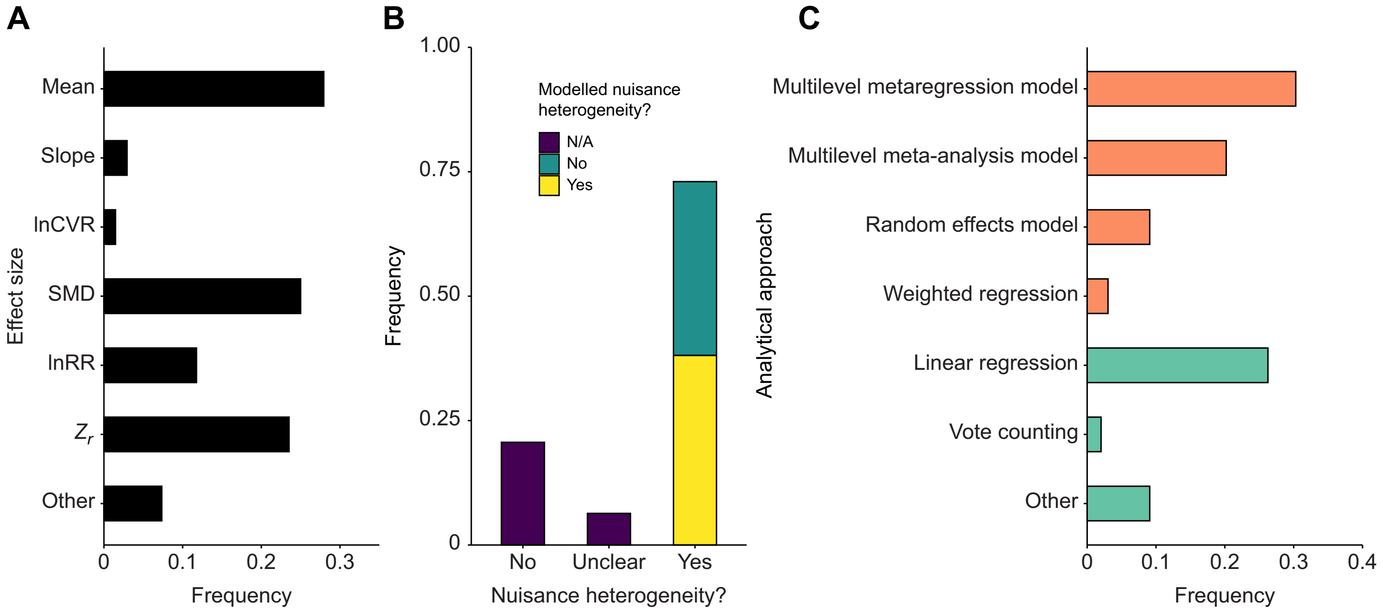

Fig. 2. Summary of meta-analytic practices in comparative physiology. (A) Different types of effect measures commonly used in meta-analyses and their relative frequency of use. (B) Frequency of studies where nuisance heterogenity had the potential to impact effect sizes (*x*-axis), and the frequencies of studies explicitly dealing with nuisance heterogeneity (bar colour). (C) Frequencies of analytical approaches used to analyse effect measures. Orange bars indicate weighted models.

within and 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 Fig. 1). As such, it is common to standardise effects by temperature. One common effect size measure already used in comparative physiology to compare physiological rates is the temperature coefficient Q_{10} (Havird et al., 2020; e.g. Rodgers et al., 2021; Seebacher et al., 2015). This effect size describes the multiplicative change in physiological rates across a 10°C temperature change. Higher Q_{10} values indicate larger changes in physiological rates. However, currently there is no formal sampling variance associated with Q_{10}, making it challenging to make use of the full power of weighted meta-analytic models. Interestingly, Q_{10} can be seen as a variant of lnRR, meaning that we can derive a Q_{10}-based effect size and sampling error using the well-known mathematical properties of lnRR. This opens up the meta-analytic toolbox and improves our ability to account for well-known sources of non-independence (Lajeunesse, 2011; Noble et al., 2017).

Prior to showing how the relevant Q_{10} 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) can be calculated as follows (but see also Senior et al., 2020):

$$\mathrm{lnRR} = \ln\left(\frac{M_1}{M_2}\right), \tag{1}$$

$$s^2_{\mathrm{lnRR}} = \left(\frac{\mathrm{SD}_1^2}{N_1 M_1^2}\right) + \left(\frac{\mathrm{SD}_2^2}{N_2 M_2^2}\right). \tag{2}$$

In Eqn 1, M_1 is the mean of group 1 (e.g. a control group), whereas M_2 is the mean of group 2 (e.g. a treatment group). The mean for group i, M_i, can be any measurement type (e.g. a physiological rate, mass, etc.) so long as the variable is measured on the ratio scale.

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	M	$\frac{\mathrm{SD}^2}{N}$	$\mathrm{CT_{min}}$, $\mathrm{CT_{max}}$, $\mathrm{LC_{50}}$, $\mathrm{LT_{50}}$, metabolic rate (MR)
log Standard deviation, lnSD	$\ln\mathrm{SD} + \frac{1}{2(N-1)}$	$\frac{1}{2(N-1)}$	Variability in $\mathrm{CT_{min}}$, $\mathrm{CT_{max}}$, MR
log Response ratio, lnRR	$\ln\left(\frac{M_1}{M_2}\right)$	$\frac{\mathrm{SD}_1^2}{M_1^2 N_1} + \frac{\mathrm{SD}_2^2}{M_2^2 N_2}$	Ratio between pollutant exposed treatment (e.g. BPA-exposed group) and control (no pollutant)
Standardised mean difference, SMD[a]	$\frac{(M_2 - M_1)}{\mathrm{SD_p}} J$	$\frac{N_1+N_2}{N_1 N_2} + \frac{\mathrm{SMD}^2}{2(N_1+N_2)}$	Difference in immune response between males and females, performance difference in the presence of stressor compared with absence of stressor
Z_r (Fisher transformation of correlation coefficient, r)	$\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$	$\frac{1}{N-3}$	Relationship between sex hormones and immune responses or metabolic rate and behaviour

N, sample size; $\mathrm{CT_{min}}$, critical thermal minimum; $\mathrm{CT_{max}}$, critical thermal maximum; $\mathrm{LC_{50}}$, median lethal concentration; $\mathrm{LT_{50}}$, median lethal temperature; BPA, bisphenol A.

[a] $J = 1 - \frac{3}{4(N_1 + N_2 - 2) - 1}$; $\mathrm{SD}_p = \sqrt{\frac{(N_1 - 1)\mathrm{SD}_1^2 + (N_2 - 1)\mathrm{SD}_2^2}{N_1 + N_2 - 2}}$.

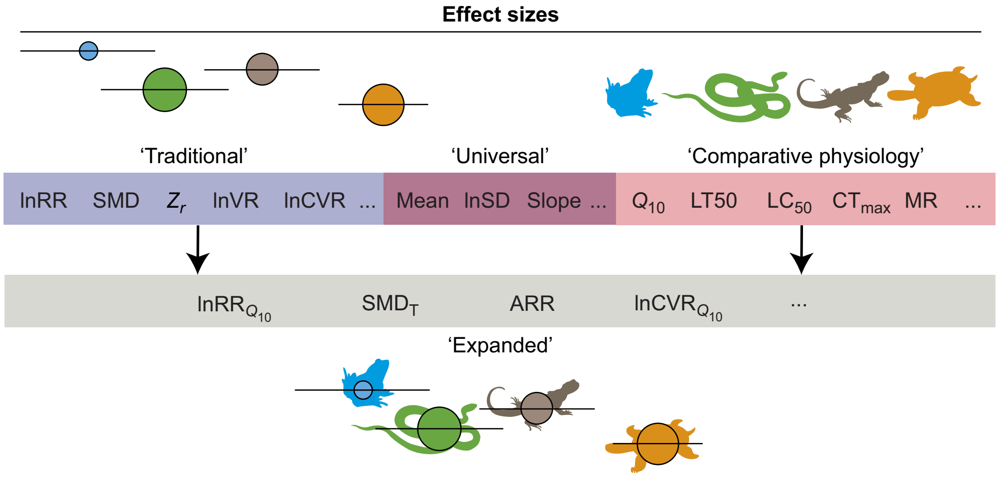

Fig. 3. Effect sizes from old to new. 'Traditional' effect sizes (top left) are somehow detached from more biologically relevant measures commonly used in 'comparative physiology' (top right), although statistical connections have always existed ('universal'; top middle). By combining traditional effects with physiological parameters, and correcting the resulting quantities for nuisance variation, we can derive a new generation of biologically relevant, easily interpretable effect measures for use in meta-analyses ('expanded'; bottom). For abbreviations of common effect sizes, see Table 1. lnRR_{Q10}, Q_{10} response ratio; lnCVR_{Q10}, log Q_{10} coefficient of variation ratio; SMD_T, temperature-corrected standardised mean difference; ARR, acclimation response ratio.

Natural log transformation of this ratio makes the effect size normally distributed (as commonly assumed by meta-analytic models). Eqn 2 is the analytical solution for the sampling variance of lnRR, where SD_1^2 and SD_2^2 are the sample standard deviations and N_1 and N_2 are the sample sizes for group 1 and 2, respectively.

The equations for lnRR and its sampling variance allow us to extend this to Q_{10}-based effect sizes. Recall that Q_{10} is described as follows:

$$Q_{10}=\left(\frac{R_2}{R_1}\right)\left(\frac{10°\text{C}}{T_2-T_1}\right). \tag{3}$$

Here, R_1 and R_2 are mean physiological rates and T_1 and T_2 are the temperatures at which these rates are measured for groups 1 and 2, respectively. Natural log transformation of Eqn 3 leads to the following log-transformed Q_{10}:

$$\text{lnRR}_{Q_{10}}=\ln\left(\frac{R_2}{R_1}\right)\left(\frac{10°\text{C}}{T_2-T_1}\right). \tag{4}$$

Eqn 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 at which these groups have been measured. Here, we will refer to this as the log Q_{10} response ratio, lnRR_{Q10}. Notably, using this effect rather than the traditional Q_{10} has two statistical advantages: (1) zero becomes biologically meaningful as zero means two rates are exactly the same, and 'significantly different from zero' means that two rates are significantly different to each other; and (2) lnRR_{Q10} is more likely to satisfy the assumption of residual normality than Q_{10} (see Hedges et al., 1999). The recognition of this equivalence means that we can calculate the sampling variance for Eqn 4 as follows:

$$s^2_{\text{lnRR}_{Q_{10}}}=\left(\frac{\text{SD}_2^2}{R_2^2N_2}+\frac{\text{SD}_1^2}{R_1^2N_1}\right)\left(\frac{10°\text{C}}{T_2-T_1}\right)^2, \tag{5}$$

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 Q_{10} analogues (and associated sampling variances) as follows:

$$\text{lnVR}_{Q_{10}}=\ln\left(\frac{\text{SD}_2}{\text{SD}_1}\right)\left(\frac{10°\text{C}}{T_2-T_1}\right), \tag{6}$$

$$s^2_{\text{lnVR}_{Q_{10}}}=\left(\frac{1}{2(N_2-1)}+\frac{1}{2(N_1-1)}\right)\left(\frac{10°\text{C}}{T_2-T_1}\right)^2. \tag{7}$$

Eqns 6 and 7 describe the change in physiological rate variance (Eqn 6) relative to 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:

$$\text{lnCVR}_{Q_{10}}=\ln\left(\frac{\text{CV}_2}{\text{CV}_1}\right)\left(\frac{10°\text{C}}{T_2-T_1}\right), \tag{8}$$

$$s^2_{\text{lnCVR}_{Q_{10}}}=\left[\frac{\text{SD}_1^2}{N_1R_1^2}+\frac{\text{SD}_2^2}{N_2R_2^2}+\frac{1}{2(N_1-1)}+\frac{1}{2(N_2-1)}\right]\left(\frac{10°\text{C}}{T_2-T_1}\right)^2, \tag{9}$$

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

Our example using Q_{10}, a well-known effect measure in comparative physiology, shows that common effect sizes can be re-formalised to account for nuisance heterogeneity (in this case, temperature). In doing so, 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 the Appendix, applying similar principles to standardise temperature differences for slopes and standardised mean differences.

Meta-analytic models to control for nuisance heterogeneity and investigate heterogeneous effects within and across studies

It will not always be possible to derive an effect size that completely controls for nuisance heterogeneity at the effect size level. 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 in relation to some mean response (e.g. quadratic relationships). 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 is the main interest in most meta-analyses anyway. Nonetheless, there are some simple statistical approaches a comparative physiologist can use to 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.

Multilevel 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, such as a 95% confidence interval (CI). Including a prediction interval (PI) can also be useful (discussed below) (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., 2017a,b). Often, meta-analyses in comparative physiology will have complex hierarchical structure. For example, multiple effect sizes may be taken from a single study on traits measured on the same individuals or from many different species that share an evolutionary history (Cinar et al., 2021 preprint; Noble et al., 2017). As such, it is far more likely that comparative physiologists will need to apply a multilevel meta-analytic model to control for the various sources of non-independence (Cinar et al., 2021 preprint; 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 drivers of effect size variance (a feature we describe below). For the purpose of our discussion, we will assume one needs to use more advanced multilevel models. Having said that, in some instances, 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 (or might not) want to apply these models (Borenstein et al., 2009; Koricheva et al., 2013; Nakagawa and Santos, 2012; Nakagawa et al., 2017a,b; Schmid et al., 2021).

Assume that we are interested in understanding the overall effect of, say, testosterone on offspring traits (e.g. see Podmokła et al., 2018). We might extract data from studies manipulating testosterone in bird eggs. Of course, these studies are diverse, spanning many species and applying different dosages of testosterone. If the studies are experimental in nature, then a standardised effect size, such as lnRR or SMD, might be applicable. We then might fit an overall multilevel meta-analytic model as follows:

$$\begin{aligned} Y_i &= \beta_0 + a_{k[i]} + sp_{k[i]} + s_{j[i]} + e_i + m_i, \\ a_{k[i]} &\sim N(0, \sigma^2_{\text{phylogeny}}\mathbf{A}), \\ sp_{k[i]} &\sim N(0, \sigma^2_{\text{species}}\mathbf{I}), \\ s_{j[i]} &\sim N(0, \sigma^2_{\text{study}}\mathbf{I}), \\ e_i &\sim N(0, \sigma^2_{\text{residual}}\mathbf{I}), \\ m_i &\sim N(0, v_i\mathbf{I}), \end{aligned} \tag{10}$$

where Y_i is the ith effect size (i=1, …, N_{ES}, the number of effect sizes), β_0 is the overall meta-analytic mean, and $a_{k[i]}$ 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 preprint). Phylogenetic effects are assumed to be normally distributed deviates sampled from a distribution with a mean of 0 and variance $\sigma^2_{\text{phylogeny}}$ [i.e. $N(0, \sigma^2_{\text{phylogeny}}\mathbf{A})$], where $\mathbf{A}$ is a phylogenetic correlation matrix derived from a phylogenetic tree. $sp_{k[i]}$ is an additional species-specific effect for the kth species applied to effect size i (i.e. these are species-specific effects because of shared ecology and other factors, Cinar et al., 2021 preprint), $s_{j[i]}$ is the study-specific effect for the jth study applied to effect size i, e_i is the effect size-specific effect (or within study effect, or residuals) for the ith effect size, and m_i is the sampling effect for the ith effect size, resulting from varying precision for each effect size, v_i (which is known as the sampling variance for the effect).

This model is a 'weighted' multilevel meta-analysis model because the estimation of the overall meta-analytic mean, β_0, is partially weighted by the inverse sampling variance of each effect size. Weighted meta-analytic models are important to use because they: (1) improve the precision on meta-analytic mean estimates by accounting for sampling variance; (2) allow for a formal analysis of heterogeneity; and (3) give precedence to higher quality studies (see below). From our survey, 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 account for sampling variance. Many studies used 'linear regression' (including linear and generalised linear mixed effects models) as the main type of model fit to the data (Fig. 2C). Of the studies using weighted meta-analytic models, 35% (N=14 meta-analyses) did not report any measure of heterogeneity, which is unfortunate given how important these measures are to interpreting effects.

Importantly, if we have effect sizes that are derived from the same sample of organisms (e.g. because traits are measured on the same sets of individuals), then we need to remove this dependency from the calculation of each v_i; we can then account for the assumed sampling covariance between effect sizes by modifying $\mathbf{I}$ to include off-diagonals that describe this covariance (see the Appendix and Noble et al., 2017). Alternatively, robust variance estimators can also be used to deal with effect size non-independence (Nakagawa et al., 2021c; Song et al., 2021). The fact that we can obtain an overall estimate of sampling variance (i.e. $\sigma^2_{\text{sampling}}$, 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., 2017a,b). This is formalised in what we call ‘heterogeneity analysis’, which we describe below.

Importantly, this model provides us with the answer to our question: across all the species and studies manipulating egg testosterone, how much of an overall effect does testosterone have on offspring traits relative to control conditions? Of course, the overall effect does not provide us with a way to interpret its magnitude in the context of how much testosterone 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 given 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, account for nuisance heterogeneity and improve effect 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:

$$Y_i = \beta_0 + \sum_{l=1}^{p} \beta_l x_{l[i]} + a_{k[i]} + sp_{k[i]} + s_{j[i]} + e_i + m_i, \qquad (11)$$

where $\sum_{l=1}^{p} \beta_l x_{l[i]}$ is simply the sum of all effects for all moderators (fixed effects; p=number of slopes), x_l. All other notation is the same as described above. Importantly, the variable x_l 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, β_0, changes. For example, assume we collected the change in *in ovo* testosterone dosage relative to a control group (i.e. in ng testosterone g^{-1} yolk) for our meta-analysis on egg testosterone effects on bird offspring traits. If the untransformed dosage, x_l, is included in the model, then the overall meta-analytic mean estimate, β_0, would be the mean effect of testosterone on traits when the dosage difference between the treatment and control group is 0 ng testosterone g^{-1} yolk (i.e. it would be reflective of unmanipulated eggs). Of course, this interpretation does not make much sense.

We can, however, apply some simple transformations to make the overall mean estimate more intuitive while also improving the interpretation of non-linear parameters and interactions estimated from the model (Gelman and Hill; Schielzeth, 2010). One simple transformation would be to centre the variable x_l by subtracting each value from the mean to create a new input variable (i.e. $c_l = x_i - \bar{x}$, where the subscript i denotes each individual effect size dosage difference). The new variable, c_l, is now centred on the mean and can replace x_l in the meta-regression model. When the model is refitted with c_l, the estimated β_0 can now be more intuitively interpreted as the overall mean effect of testosterone on offspring traits at an average dosage difference between treatment and control groups. Importantly, by centring, we retain the original units (ng testosterone g^{-1} yolk) of the variable, which can be very useful to ease the interpretation of the effect magnitude (e.g. when the dosage difference is 10 ng testosterone g^{-1} yolk).

Given it is very common to estimate and compare overall meta-analytic means (i.e. β_0) in meta-analyses, centring can be particularly useful in making these means more interpretable. Mean centring can even be extended to include other ‘nuisance’ variables that might creep into the dataset; the same centring approach can be done with a second variable, and the overall mean adjusted and interpreted in the context of this new variable along with dosage. Of course, if one is not interested in the magnitude of the effect, we can 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, such as temperature, can then be modelled (accounting for sampling variance) to understand mean trait 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).

Centring 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 fitted without compromising interpretations, such as when one includes non-linear parameters (e.g. higher-order polynomials) and interactions with other variables. In other words, linear or main effects can still be interpreted normally in the presence of non-linear effects or interactions fit in the model (Gelman and Hill; Schielzeth, 2010). We provide examples in the online supplementary material (https://daniel1noble.github.io/nuisance_heterogeneity/) along with R code to show readers how the models we describe above can be fitted, and how meta-analytic means can be adjusted for treatment heterogeneity.

Heterogeneity analysis: how much nuisance heterogeneity actually exists?

As we have already emphasised, quantifying effect size heterogeneity provides the impetus for exploring why effects might vary within and across studies, while also providing context for interpreting the generality of overall effects. There are several measures commonly used to quantify effect size heterogeneity (Borenstein, 2019); however, in ecological and evolutionary meta-analysis, heterogeneity is often calculated and reported as I^2_{total} as follows (if using a multilevel meta-analytic model as we describe above):

$$I^2_{\text{total}} = \frac{\sigma^2_{\text{study}} + \sigma^2_{\text{phylogeny}} + \sigma^2_{\text{species}} + \sigma^2_{\text{residual}}}{\sigma^2_{\text{study}} + \sigma^2_{\text{phylogeny}} + \sigma^2_{\text{species}} + \sigma^2_{\text{residual}} + \sigma^2_{\text{sampling}}}, \qquad (12)$$

where $\sigma^2_{\text{total}} = \sigma^2_{\text{study}} + \sigma^2_{\text{phylogeny}} + \sigma^2_{\text{species}} + \sigma^2_{\text{residual}} + \sigma^2_{\text{sampling}}$ is the total effect size variance and $\sigma^2_{\text{sampling}}$ is the ‘typical’ sampling error variance calculated as:

$$\sigma^2_{\text{sampling}} = \sum_{i}^{N_{ES}} w_i(k-1) \Big/ \left[\left(\sum w_i \right)^2 + \sum w_i^2 \right], \qquad (13)$$

where k is the number of studies and the weights, w_i=1/v_i, can be calculated using the inverse of the sampling variance (v_i) for each effect size i.

Quantifying heterogeneity using I^2_{total} provides a formal way to assess the relative amount of variation that is the result of ‘true’ biological or methodological differences as opposed to chance (i.e. sampling variance) (Higgins and Thompson, 2002; Nakagawa and Santos, 2012). In other words, I^2_{total} is the percentage of variance between effect sizes after removing the effects of sampling error (Higgins and Thompson, 2002). Note that 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 within and across studies. Some of the variation is clearly the result of working with different species or strains, or measuring different traits and outcome measures. However, some of this variation also arises because of methodological differences across studies (e.g. different temperature or dosage 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 producing various I^2 estimates that 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): $I^2_{\text{study}} = \sigma^2_{\text{study}}/\sigma^2_{\text{total}}$, where σ^2_{study} is the study-specific variance.

Some readers may notice the similarities between I^2 and R^2 (Nakagawa and Schielzeth, 2010, 2013; Nakagawa et al., 2017a,b). If we want to formally assess how much effect size variation is the result of nuisance heterogeneity (e.g. 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):

$$R^2_{\text{marginal}} = \frac{\sigma^2_{\text{fixed}}}{\sigma^2_{\text{fixed}} + \sigma^2_{\text{study}} + \sigma^2_{\text{phylogeny}} + \sigma^2_{\text{species}} + \sigma^2_{\text{residual}}}. \quad (14)$$

Note that this formula does not include $\sigma^2_{\text{sampling}}$, 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, R^2_{marginal} provides a formal means to assess just how much variation in effects is the result of nuisance heterogeneity. Other biological moderators of interest could also be included as fixed effects and R^2_{marginal} can be 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 variation at the study or species level (e.g. reproductive mode, endothermy, etc.). R^2_{marginal}, or how we like to think of it in this context, R^2_{nuisance}, may be 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, dosage 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 I^2 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). Prediction intervals describe the range of plausible effect-size values expected from a future study (Nakagawa et al., 2021b). This is different from a confidence interval that expresses the range of uncertainty around a statistical parameter estimate. Unlike I^2, prediction intervals (PIs) are probably more meaningful in the context of meta-analysis because they explicitly incorporate measures of dispersion around a mean effect size. They provide 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 might collect experimental studies measuring fish swimming performance under high salinity treatments (~15–20 ppm NaCl) compared with freshwater (~0 ppm NaCl) controls, using SMD to quantify the effect salinity had on swim performance. We conducted a meta-analysis and estimated an overall SMD of 0.50 with a 95% 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 of 0.85, depending on the population. In addition, if we were to repeat a similar 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). We encourage more meta-analysts to report these (for examples of PIs, see supplementary material at https://daniel1noble.github.io/nuisance_heterogeneity/).

Conclusions

Comparative physiologists are interested in meta-analysing effects across a diversity of species, experimental designs and environments. Importantly, sampling variances for many effect sizes commonly used by comparative physiologists (e.g. CT_{max}, Q_{10}, etc.) can be formalized and powerful meta-analytic models can be easily applied to these effect measures (Fig. 3). Although 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, factors such as temperature and dosage differences across studies may generate obvious sources of heterogeneity that may not be of prime interest. Instead, comparative physiologists might simply want to focus on biological drivers of effect-size variability and want an effect size or analytical approach that allows them to remove these 'nuisances'. Surveying the comparative physiology literature shows that nuisance heterogenity is common, and is often not completely dealt with in a meta-analysis. Here, we provide a set of tools to deal with nuisance heterogeneity (Fig. 1) by reinventing standard effect sizes and/or using mean centred nuisance variables in meta-regression models. Estimating complex meta-regression models with a multitude of fixed effects might result, however, in more challenging model interpretation. As such, a combined approach might be more desirable. Nuisance heterogeneity can be dealt with by using an appropriate effect size and associated sampling variance in combination with meta-regression models that contain moderators that test biological hypotheses of interest. Indeed, this may even be a necessity to simplify the model. The set of tools we describe provide clarification around the importance and limitations of heterogeneity. We also hope to provide new ways to help comparative physiologists communicate effect measures more richly to guide our understanding and decisions around pressing global challenges.

APPENDIX

Examples of how to derive sampling variances

Here, we show how to obtain sampling variance for a slope (or a 'rate of change') when you have measurements of a physiological trait at two points along an environmental gradient (e.g. temperature, salinity, pH). We then derive sampling variance for the difference between two slopes, and demonstrate how SMD can be corrected to account for differences in both units (e.g. cm or mg) and points on an environmental gradient (e.g. temperature of 24°C and 30°C). Finally, we introduce a useful approximation technique known as 'the Delta method'. The sampling variances associated with these 'new' effect sizes will allow comparative physiologists to take advantage of powerful meta-analytic models.

Sampling variance for a slope between two points

Let us start with a real example that comparative physiologists can easily relate with, the slope of responses to temperature acclimation, or as we call it, the acclimation response ratio (ARR) (Pottier et al., 2021). ARR can be defined as a slope for acclimated physiological responses at two different temperature points, defined as:

$$\mathrm{ARR} = \frac{M_1 - M_2}{T_2 - T_1}, \tag{A1}$$

where T is temperature (°C) and $T_2 > T_1$, and M_1 and M_2 are the average physiological responses (e.g. mean $\mathrm{CT_{max}}$) at temperature points T_1 and T_2, respectively. Many studies might manipulate multiple variables using a fully factorial design (e.g. temperature and pH); however, for simplicity we assume the study only manipulates temperature. In the supplementary information (https://daniel1noble.github.io/nuisance_heterogeneity/), we also show how to derive ARR from fully factorial studies (i.e. main effects and interactions).

To obtain the sampling variance for this equation (slope), we first need to describe some basic properties of variance. Let us assume M_1 is a random variable drawn from a distribution that can be characterised by a mean and standard deviation (note that this standard deviation is not the 'sample' but the 'sampling' standard deviation, which is often referred to as standard error; see fig. 1 in Nakagawa et al., 2021a). Multiplying it by a constant (a) will change the variance by the square of that constant (a^2) while adding or subtracting the constant (b) does not change the variance of M_1. This can be summarized as:

$$\sigma^2(aM_1 \pm b) = a^2\sigma^2_{M_1}. \tag{A2}$$

Also, when adding two random variables (M_1 and M_2), the combined variance is the sum of the variance of M_1 and the variance of M_2 plus 2 times the covariance between M_1 and M_2. This relationship can be written as:

$$\begin{aligned}\sigma^2(M_1 \pm M_2) &= \sigma^2_{M_1} + \sigma^2_{M_2} \pm 2\mathrm{Cov}(M_1, M_2)\\ &= \sigma^2_{M_1} + \sigma^2_{M_2} \pm 2\mathrm{Cor}(M_1, M_2)\sqrt{\sigma^2_{M_1}\sigma^2_{M_2}},\end{aligned} \tag{A3}$$

where the covariance $\mathrm{Cov}(M_1,M_2)$ equals the correlation multiplied by the square root of the product of two variances $2\mathrm{Cor}(M_1, M_2)\sqrt{\sigma^2_{M_1}\sigma^2_{M_2}}$.

Importantly, when M_1 and M_2 are independent of each other, their covariance is 0. In other words, if measurements are taken from two different groups of animals at two different temperatures (T_1 and T_2), then the covariances between these two sets of measurements are 0.

Therefore, when M_1 and M_2 are independent, we can obtain the sampling variance for ARR as:

$$\sigma^2_{\mathrm{ARR}} = \left(\frac{1}{T_2 - T_1}\right)^2\left(\frac{\mathrm{SD}_1^2}{N_1} + \frac{\mathrm{SD}_2^2}{N_2}\right), \tag{A4}$$

where SD_1 and SD_2 and N_1 and N_2 are standard deviations and sampling sizes at temperatures T_1 and T_2, respectively. Readers may find it difficult to see how we obtained this equation. Let us explain further. The (sampling) standard error for M_1 is $\mathrm{SE}_1 = \mathrm{SD}_1/\sqrt{N_1}$. Given this, the sampling variance for M_1 is SD_1^2/N_1, and the sampling variance for M_2 is SD_2^2/N_2. You can see those elements of the sampling variance of ARR in the formula above. The term $[1/(T_2–T_1)]^2$ comes from recognizing $1/(T_2–T_1)$ as a constant.

In ecological and evolutionary studies, it is not uncommon that we measure the same group of organisms at two temperatures. However, if we do, then we need to add the covariance between M_1 and M_2 [$\mathrm{Cov}(M_1,M_2)$] in the equation. Note that, as above, the covariance equals the correlation multiplied by the square root of two (sampling) variances. Therefore, the sampling variance of ARR can be now written as:

$$\sigma^2_{\mathrm{AAR}} = \left(\frac{1}{T_2 - T_1}\right)^2\left(\frac{\mathrm{SD}_1^2}{N_1} + \frac{\mathrm{SD}_2^2}{N_2} - 2r\sqrt{\frac{\mathrm{SD}_1^2}{N_1}\frac{\mathrm{SD}_2^2}{N_2}}\right). \tag{A5}$$

By assuming the numbers of organisms (N) are the same at the two temperatures, we can slightly simplify this formula:

$$\sigma^2_{\mathrm{AAR}} = \left(\frac{1}{T_2 - T_1}\right)^2\left(\frac{\mathrm{SD}_1^2 + \mathrm{SD}_2^2 - 2r\mathrm{SD}_1\mathrm{SD}_2}{N}\right), \tag{A6}$$

where r is the correlation between a set of measurements M_1 and M_2 from the same individuals at two points T_1 and T_2. As you may notice, we need the raw data to calculate r. Therefore, in reality, we often need to assume a certain value of r. When we do not have an estimate of r we can reasonably assume it to be 0.5 (Noble et al., 2017).

Sampling variance for the difference between two slopes

Now let us assume that we want to know the difference between two different ARR values: e.g. female ARR ($\mathrm{ARR_f}$) and male ARR ($\mathrm{ARR_m}$). Such a difference (ARRD) can be written as:

$$\mathrm{ARRD} = \mathrm{ARR_f} - \mathrm{ARR_m}. \tag{A7}$$

Using the properties of variance and the equations from above, the sampling variance for ARRD can be derived; when measurements at two temperature points are independent, as:

$$\sigma^2_{\mathrm{ARRD}} = \left(\frac{1}{T_2 - T_1}\right)^2\left(\frac{\mathrm{SD}_{\mathrm{f1}}^2}{N_{\mathrm{f1}}} + \frac{\mathrm{SD}_{\mathrm{f2}}^2}{N_{\mathrm{f2}}} + \frac{\mathrm{SD}_{\mathrm{m1}}^2}{N_{\mathrm{m1}}} + \frac{\mathrm{SD}_{\mathrm{m2}}^2}{N_{\mathrm{m2}}}\right), \tag{A8}$$

where subscripts f and m stand for females and males, respectively.

Similarly, the dependent version of the sampling variance can be written as:

$$\sigma^2_{\mathrm{ARRD}} = \left(\frac{1}{T_2 - T_1}\right)^2 \left(\frac{\mathrm{SD}_{\mathrm{f1}}^2 + \mathrm{SD}_{\mathrm{f2}}^2 - 2r\mathrm{SD}_{\mathrm{f1}}\mathrm{SD}_{\mathrm{f2}}}{N_{\mathrm{f}}} + \frac{\mathrm{SD}_{\mathrm{m1}}^2 + \mathrm{SD}_{\mathrm{m2}}^2 - 2r\mathrm{SD}_{\mathrm{m1}}\mathrm{SD}_{\mathrm{m2}}}{N_{\mathrm{m}}}\right). \tag{A9}$$

Here, we have derived sampling variances for 'temperature' ARR and ARRD for both independent and dependent cases. These formulas can also be applied to changes in other factors (e.g. salinity, pH, oxygen) by changing the temperature constant. Also, our calculation can easily account for methodological inconsistencies. For example, it may be possible that males and females were measured at slightly different temperatures; say females at T_3 and T_4 (and $T_4 > T_3$) and males at T_5 and T_6 (and

$T_6 > T_5$), like below:

$$\sigma^2_{\mathrm{ARRD}} = \left(\frac{1}{T_4 - T_3}\right)^2 \left(\frac{\mathrm{SD}^2_{\mathrm{f3}} + \mathrm{SD}^2_{\mathrm{f4}} - 2r\mathrm{SD}_{\mathrm{f3}}\mathrm{SD}_{\mathrm{f4}}}{N_{\mathrm{f}}}\right) + \left(\frac{1}{T_6 - T_5}\right)^2 \left(\frac{\mathrm{SD}^2_{\mathrm{m5}} + \mathrm{SD}^2_{\mathrm{m6}} - 2r\mathrm{SD}_{\mathrm{m5}}\mathrm{SD}_{\mathrm{m6}}}{N_{\mathrm{m}}}\right). \tag{A10}$$

Also, note that the difference between slopes does not need to be that of males and females. This formula can be used to compare any two treatments or biological groups.

Controlling for unit differences and temperature across studies

Calculating ARR assumes that the units for the mean difference across studies is consistent. Frequently, this will not be the case for meta-analyses that hope to synthesise a wide array of traits that vary in their units (e.g. µl g^{-1}, g, s, min). Fortunately, lnRR and SMD (Table 1) already control for unitary differences. As such, we can apply similar logic that we discuss above to correct SMD for both temperature and unitary differences. Table 1 provides the formula for SMD, its pooled standard deviation (SD_p) and sampling variance. Just like with ARR, we can apply our temperature correction to the SMD formula as follows:

$$\mathrm{SMD}_{\mathrm{T}} = \frac{(M_1 - M_2)}{\mathrm{SD}_{\mathrm{p}}(T_2 - T_1)} J. \tag{A11}$$

Here, we can see that the difference between M_1 and M_2 is standardised by the pooled SD, correcting for differences in measurement units. Additionally, dividing SMD by the temperature difference results in further correcting the effect size by the applied temperature difference. J is a small sample correction (see Table 1). To derive the sampling variance for SMD_T, we can apply the same principle as we did above to derive a sampling variance as:

$$s^2_{\mathrm{SMD}_{\mathrm{T}}} = \left(\frac{N_1 + N_2}{N_1 N_2} + \frac{\mathrm{SMD}^2_{\mathrm{T}}}{2(N_1 + N_2)}\right)\left(\frac{1}{T_2 - T_1}\right)^2. \tag{A12}$$

The Delta method

Deriving sampling variances using common properties of variance has its limits. Therefore, we introduce a practical and widely applicable method to obtain approximate variance when the basic properties of variance cannot be applied. This is where the Delta method comes in. The general form of this method can be written as:

$$\mathrm{Var}(f(X)) \approx \mathrm{Var}(X) \times (f'(X))^2, \tag{A13}$$

where $f(X)$ represents the function of the random variable X, and importantly, $f'(X)$ is the first derivative of $f(X)$. Let us demonstrate this with a concrete example by re-deriving the sampling variance for lnRR defined as:

$$\mathrm{lnRR} = \ln\left(\frac{M_1}{M_2}\right) = \ln M_1 - \ln M_2. \tag{A14}$$

Here, $f(M_1)=\ln M_1$. By applying $f'(M_1)=1/M_1$ (i.e. the first derivative of $\ln M_1$ is $1/M_1$) to the Delta method and using the variance's basic properties, we have:

$$\sigma^2_{\mathrm{lnRR}} \approx \left(\frac{\mathrm{SD}^2_1}{N_1}\right)\left(\frac{1}{M_1}\right)^2 + \left(\frac{\mathrm{SD}^2_2}{N_2}\right)\left(\frac{1}{M_2}\right)^2 = \left(\frac{\mathrm{SD}^2_1}{N_1 M^2_1}\right) + \left(\frac{\mathrm{SD}^2_2}{N_2 M^2_2}\right). \tag{A15}$$

Note that the first term, SD^2_1/N_1 is once again the sampling variance for group 1, and we simply multiply this term by the square of the first derivative of M_1 as described by the Delta method. Readers may notice that this formula is now equivalent to the sampling variance for lnRR (Table 1). We can extend this sampling variance to include dependency between the groups as:

$$\sigma^2_{\mathrm{lnRR}} \approx \left(\frac{\mathrm{SD}^2_1}{N_1 M^2_1}\right) + \left(\frac{\mathrm{SD}^2_2}{N_2 M^2_2}\right) - 2r\sqrt{\frac{\mathrm{SD}^2_1}{N_1 M^2_1}}\sqrt{\frac{\mathrm{SD}^2_2}{N_2 M^2_2}}, \tag{A16}$$

where r is the correlation between $\ln M_1$ and $\ln M_2$ when organisms are measured multiple times (with r often assumed to be 0.5), while r is 0 when two sets of measurements are independent.

Finally, we note that, by using the basic properties of variance and the Delta method, one can derive sampling variance for most effect size measures. For instance, these methods allowed us to derive sampling variance for our 'new' effect sizes (e.g. $\ln Q_{10}$; see 'Comparing changes in mean physiological rates, Q_{10}') and were used in Nakagawa et al. (2015) and Senior et al. (2020). We also refer the reader to Nakagawa et al. (2017a) for additional details on the Delta method.

Acknowledgements
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 thank Megan Head and two anonymous reviewers for very helpful comments on previous drafts.

Competing interests
The authors declare no competing or financial interests.

Author contributions
Conceptualization: D.W.A.N., S.N.; Methodology: D.W.A.N., S.N.; Software: D.W.A.N., S.N., M.L., P.P., R.E.O., S.D., S.B.; Validation: M.L., P.P., R.E.O., S.D., S.B.; Formal analysis: D.W.A.N., S.N., M.L., P.P., R.E.O., S.D., S.B.; Resources: D.W.A.N., S.N.; Data curation: M.L., P.P., R.E.O., S.D., S.B.; Writing - original draft: D.W.A.N., S.N.; Writing - review & editing: M.L., P.P., R.E.O., S.D., S.B.; Visualization: D.W.A.N., S.N., M.L., P.P., R.E.O., S.D., S.B.; Project administration: D.W.A.N., S.N.; Funding acquisition: D.W.A.N., S.N.

Funding
D.W.A.N., S.M.D., S.N., M.L. and R.E.O. were supported by Australian Research Council Discovery Projects (DP210101152 to D.W.A.N., DE180100202 to S.M.D., and DP200100367 to S.N., M.L. and D.W.A.N.). P.P. and S.B. were supported by a University of New South Wales Scientia PhD Scholarship. This work was supported by the Australian National University.

Data availability
Supplementary information, data, code and results associated with this paper can all be found at: https://daniel1noble.github.io/nuisance_heterogeneity/.

References

Arnqvist, G. and Wooster, D. (1995). Meta-analysis: Synthesizing research findings in ecology and evolution. *Trends Ecol. Evol.* **10**, 236-240. doi:10.1016/S0169-5347(00)89073-4

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

Borenstein, M., Hedges, L. V., Higgens, J. P. T. and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. West Sussex, UK: John Wily & Sons, Ltd.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. and Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365-376. doi:10.1038/nrn3475

Carpenter, C. J. (2020). Meta-analyzing apples and oranges: How to make applesauce instead of fruit salad. *Hum. Commun. Res.* **46**, 322-333. doi:10.1093/hcr/hqz018

Cinar, O., Nakagawa, S. and Viechtbauer, W. (2021). Phylogenetic multilevel meta-analysis: a simulation study on the importance1of modeling the phylogeny. *EcoEvoRxiv*. doi:10.32942/osf.io/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. H. Cooper, L. V. Hedges and J. C. Valentine), 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. *Method. Ecol. Evol.* **12**, 1705-1720. doi:10.1111/2041-210X.13654

Forstmeier, W., Wagenmakers, E. J. and Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biol. Rev.* **92**, 1941-1968. doi:10.1111/brv.12315

Gallo, P. S. (1978). Meta-analysis: A mixed meta-phor? *Am. Psychol.* **33**, 515-517. doi:10.1037/0003-066X.33.5.515

Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierachical Models*. Cambridge, UK: Cambridge University Press.

Glass, G. V. (2015). Meta-analysis at middle age: a personal history. *Res. Synth. Methods* **6**, 221-231. doi:10.1002/jrsm.1133

Gurevitch, J. and Hedges, L. V. (1999). Statistical issues in ecological meta-analyses. *Ecology* **80**, 1142-1149. doi:10.1890/0012-9658(1999)080[1142:SIIEMA]2.0.CO;2

Gurevitch, J., Koricheva, J., Nakagawa, S. and Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature* **555**, 176-182. doi:10.1038/nature25753

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. *Funct. Ecol.* **34**, 1015-1028. doi:10.1111/1365-2435.13534

Hedges, L. V., Gurevitch, J. and Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology* **80**, 1150-1156. doi:10.1890/0012-9658(1999)080[1150:TMAORR]2.0.CO;2

Higgins, J. P. T. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539-1558. doi:10.1002/sim.1186

Holtmann, B., Lagisz, M. and Nakagawa, S. (2016). Metabolic rates, and not hormone levels, are a likely mediator of between-individual differences in behaviour: a meta-analysis. *Funct. Ecol.* **31**, 685-696. doi:10.1111/1365-2435.12779

Jennions, M. D. and Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav. Ecol.* **14**, 438-445. doi:10.1093/beheco/14.3.438

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

Koricheva, J., Gurevitch, J. and Mengersen, K. (2013). *Handbook of Meta-Analysis in Ecology and Evolution*. Princeton, New Jersey: Princeton University Press.

Lajeunesse, M. J. (2010). Achieving synthesis with meta-analysis by combining and comparing all available studies. *Ecology* **91**, 2561-2564. doi:10.1890/09-1530.1

Lajeunesse, M. J. (2011). On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology* **92**, 2049-2055. doi:10.1890/11-0423.1

Lajeunesse, M. J. (2015). Bias and correction for the log response ratio in ecological meta-analysis. *Ecology* **96**, 2056-2063. doi:10.1890/14-2402.1

Morrissey, M. B. (2016). Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *J. Evol. Biol.* **29**, 1862-1904.

Nakagawa, S. and Santos, E. S. (2012). Methodological issues and advances in biological meta-analysis. *Evol. Ecol.* **26**, 1253-1274. doi:10.1007/s10682-012-9555-5

Nakagawa, S. and Schielzeth, H. (2010). Repeatbility for gaussian and non-gaussian data: a practical guide for biologists. *Biol. Rev.* **85**, 935-956. doi:10.1111/j.1469-185X.2010.00141.x

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalised linear mixed effects models. *Method. Ecol. Evol.* **4**, 133-142. doi:10.1111/j.2041-210x.2012.00261.x

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. *Method. Ecol. Evol.* **6**, 143-152. doi:10.1111/2041-210X.12309

Nakagawa, S., Johnson, P. C. and Schielzeth, H. (2017a). The coefficient of determination r 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* **14**, 20170213. doi:10.1098/rsif.2017.0213

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

Nakagawa, S., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D. W. A., Parkar, T. H., Sánchez-Tójar, A., Yang, Y. and O'Dea, R. E. (2021a). Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods Ecol. Evol.* doi:10.1111/2041-210X.13724

Nakagawa, S., Lagisz, M., O'Dea, R. E., Rutkowska, J., Yang, Y., Noble, D. W. A. and Senior, A. M. (2021b). The orchard plot: Cultivating forest plots for use in ecology, evolution and beyond. *Res. Synth. Method.* **12**, 4-12. doi:10.1002/jrsm.1424

Nakagawa, S., Senior, A. M., Viechtbauer, W. and Noble, D. W. A. (2021c). An assessment of statistical methods for non-independent data in ecological meta-analyses: comment. *Ecology*, e03490. doi:10.1002/ecy.3490

Noble, D. W. A., Lagisz, M., O'Dea, R. E. and Nakagawa, S. (2017). Non-independence and sensitivity analyses in ecological and evolutionary meta-analyses. *Mol. Ecol.* **26**, 2410-2425. doi:10.1111/mec.14031

Noble, D. W. A., Stenhouse, V. and Schwanz, L. E. (2018). Developmental temperatures and phenotypic plasticity in reptiles: A systematic review and meta-analysis. *Biol. Rev.* **93**, 72-79. doi:10.1111/brv.12333

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. *Biol. Rev.* **96**, 1695-1722. doi:10.1111/brv.12721

Podmokła, E., Drobniak, S. M., Rutkowska, J. (2018). Chicken or egg? Outcomes of experimental manipulations of maternally transmitted hormones depend on administration method – a meta-analysis. *Biol. Rev.* **93**, 1499-1517. doi:10.1111/brv.12406

Pottier, P., Burke, S., Drobniak, S. M., Lagisz, M. and Nakagawa, S. (2021). Sexual (in) equality? A meta-analysis of sex differences in thermal acclimation capacity across ectotherms. *Funct. Ecol.* **35**, 1015-1028. doi:10.1111/1365-2435.13899.

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. *J. Exp. Biol.* **224**, 1-12. doi:10.1242/jeb.228213

Rothstein, H. R., Sutton, A. J. and Borenstein, M. (2005). Publication bias in meta-analysis: prevention, assessment and adjustments, pp. 1-376. Chichester: Wiley.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Method. Ecol. Evol.* **1**, 103-113. doi:10.1111/j.2041-210X.2010.00012.x

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. *Nat. Clim. Chang.* **5**, 61-66. doi:10.1038/nclimate2457

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

Senior, A. M., Viechtbauer, W. and Nakagawa, S. (2020). Revisiting and expanding the meta-analysis of variation: the log coefficient of variation ratio. *Res. Synth. Methods* **11**, 553-567. doi:10.1002/jrsm.1423

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. doi:10.1002/ecy.3578

Stewart, G. B. (2010). Meta-analysis in applied ecology. *Biol. Lett.* **6**, 78-81. doi:10.1098/rsbl.2009.0546

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. *Proc. Natl. Acad. Sci. USA* **111**, 5610-5615. doi:10.1073/pnas.1316145111

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. *Am. Nat.* **190**, 185-199. doi:10.1086/692326

White, C. R., Phillips, N. F. and Seymour, R. S. (2006). The scaling and temperature dependence of vertebrate metabolism. *Biol. Lett.* **2**, 125-127. doi:10.1098/rsbl.2005.0378

Wu, N. C. and Seebacher, F. (2020). Effect of the plastic pollutant bisphenol a on the biology of aquatic organisms: a meta-analysis. *Glob. Change Biol.* **26**, 3821-3833. doi:10.1111/gcb.15127