Title

Full title: A pluralistic framework for measuring and stratifying heterogeneity in meta-analyses

Short title: Measuring and stratifying for heterogeneity

Authors

Yefeng Yang[1], Daniel W. A. Noble[2], Rebecca Spake[3], Alistair M. Senior[4], Malgorzata Lagisz[1] †, Shinichi Nakagawa[1,5*] †

Affiliations

[1]Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia

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

[3]Ecology and Evolutionary Biology Research Division, School of Biological Sciences, University of Reading, RG6 6EX, Reading, UK

[4]Charles Perkins Centre, Sydney Precision Data Science Centre, School of Life and Environmental Sciences and School of Mathematics and Statistics, The University of Sydney, Sydney, NSW 2006, Australia

[5]Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, 904-0495, Japan

*Correspondence

E-mail: s.nakagawa@unsw.edu.au (SN), yefeng.yang1@unsw.edu.au (YY)

† Equal senior author

Abstract

1. Measuring heterogeneity, or inconsistency, among effect sizes is a crucial step for interpreting the meta-analytic evidence across diverse taxonomic groups and spatiotemporal contexts. However, ecologists and evolutionary biologists often interpret mean population effects (i.e., meta-analytic mean effect size) as consistent, either explicitly or implicitly, without proper heterogeneity quantification, thus assuming consistency in effects across contexts.

2. Here, we present a pluralistic approach aimed at quantifying heterogeneity by introducing complementary measures, each of which decomposes (stratifies) heterogeneity into within- and between-study variances. These measures include the traditional I^2, stratified I^2, the newly derived coefficient of variation (CV), and its transformation (M).

3. To demonstrate the benefits of the combined use of these measures, we synthesize 512 ecological and evolutionary meta-analyses. We show that total heterogeneity (variance in true effects) is, on average, ten times larger than statistical noise (sampling variance), contributing to 91% of the observed variance (median I^2 = 91%). This amount of heterogeneity is nearly twice the size of the mean population effect (median CV = 1.8 and transformation M = 0.6), indicating substantial variation among studies within a meta-analysis.

4. Surprisingly, despite a high amount of total heterogeneity is present in most meta-analyses, half of the meta-analyses had low among-study variance (and high within-study variance), indicating the meta-analytic mean effect could be generalizable across studies.

5. Our meta-synthesis can serve as new benchmarks for the interpretation of heterogeneity. Our proposed pluralistic approach provides our recommendations on how to quantify and report heterogeneity. Collectively, we could accelerate the future quest for generalizability of ecological and evolutionary phenomena via a better understanding of meta-analytic heterogeneity.

Introduction

Meta-analytic modelling is widely used to test ecological and evolutionary hypotheses and informing conservation and environmental policies (Gurevitch *et al.* 2018). This feat is accomplished through one or more of three procedures. Firstly, meta-analysis quantitatively estimates the mean population effect (meta-analytic mean effect size) across effect sizes sampled from different contexts (Nakagawa & Santos 2012; Noble *et al.* 2022; Yang *et al.* 2022), characterising the central tendency of a focal ecological and evolutionary effect. Secondly, effect modifiers or moderators explaining variation in effect sizes are identified (context-specific effects; Nakagawa & Santos 2012). Third, meta-analysis can quantify variability in study outcomes, the "heterogeneity" among effect sizes. Heterogeneity helps indicate the degree of inconsistency or 'context dependence' of study findings, with high heterogeneity indicating high variability among effect sizes that underpin the mean population effect. . High heterogeneity thus precludes generality of the mean effect size, and signals a need to further identify the drivers of effect size variation. Without quantifying heterogeneity, it is difficult to interpret both the overall trends and context-specific effects (Senior *et al.* 2016).

While meta-analyses of a collection of studies using similar protocols for single species have clear interpretations, the interpretation of average population effects across diverse taxonomic groups and spatiotemporal contexts can be problematic. However, ecologists and evolutionary biologists often either explicitly or implicitly interpret the mean population effect and context-specific effects as consistent across contexts (Spake *et al.* 2022), and thus transferable to a broad, largely unspecified target context. The mean population effect size is only generalizable and transferable across the contexts when the meta-analytic evidence pool does not respond to effect modifiers, leading to low amount of the variability around the true effect size (i.e., low heterogeneity). Until now, the significance of heterogeneity in interpreting meta-analytic evidence has been largely overlooked. Indeed, surveys have revealed that heterogeneity statistics are not routinely reported (Senior *et al.* 2016; Yang *et al.* 2022; Nakagawa *et al.* 2023).

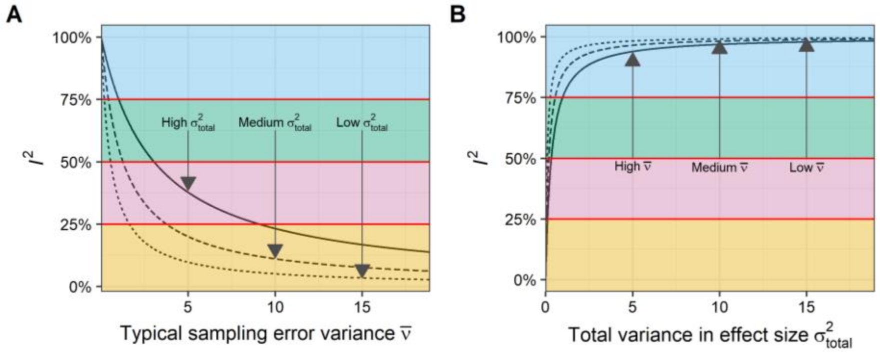

Fig. 1:

The interpretation of total I^2 can be ambiguous and can lead to incorrect conclusions about the magnitude of heterogeneity. (A) A large estimated total I^2 value could be due to small typical sampling error variances $\bar{v}$ (i.e., low statistical noise; Equation 3). (B) On the other hand, a large total I^2 value could also result from a large true heterogeneity. Values of σ^2_{total} and $\bar{v}$ were derived from their empirical distributions based on 512 meta-analyses (see Figs. S1 and S2). Total I^2 values were calculated using Equations 2 and 3. High, medium, and low σ^2_{total} (and $\bar{v}$) denote the 25%, 50%, and 75% percentiles of their empirical distributions (Table 1). Three horizontal lines denote the conventional thresholds for the use of I^2 to interpret the magnitude of heterogeneity (Higgins *et al.* 2003).

Currently, measuring and interpreting meta-analytic heterogeneity is challenging for two major reasons. First, no single heterogeneity metric provides a holistic interpretation of generalizability (Cairns & Prendergast 2022). Currently, the I^2 statistic is a popular metric that quantifies the proportion of variance due to differences between effect sizes rather than by statistical noise (i.e., sampling variance; Higgins & Thompson 2002; Rücker *et al.* 2008). The biological interpretation of I^2, however, is ambiguous (IntHout *et al.* 2016) because a small absolute heterogeneity can lead to a high I^2 due to small statistical noise (see Fig. 1; Rücker *et al.* 2008; IntHout *et al.* 2016; Borenstein *et al.* 2017). Second, meta-analyses typically focus on estimating total heterogeneity only (Nakagawa & Santos 2012), despite the hierarchical nature of real biological data structures (Noble *et al.* 2022; Nakagawa *et al.* 2023). Explicitly decomposing effect size heterogeneity across hierarchical levels (i.e., stratification) enables a more nuanced configurative account of the meta-analytic evidence, and helps identify contextual factors (Nakagawa & Santos 2012) that drive context dependence. For example, in a multi-taxon meta-analysis, if stratification of studies by species yields low heterogeneity at the taxon level, the focal effect still can be generalizable across taxon (in terms of accounting for within-taxon variation; Fig. 2). This is so, even if the total heterogeneity remains high (Senior *et al.* 2016).

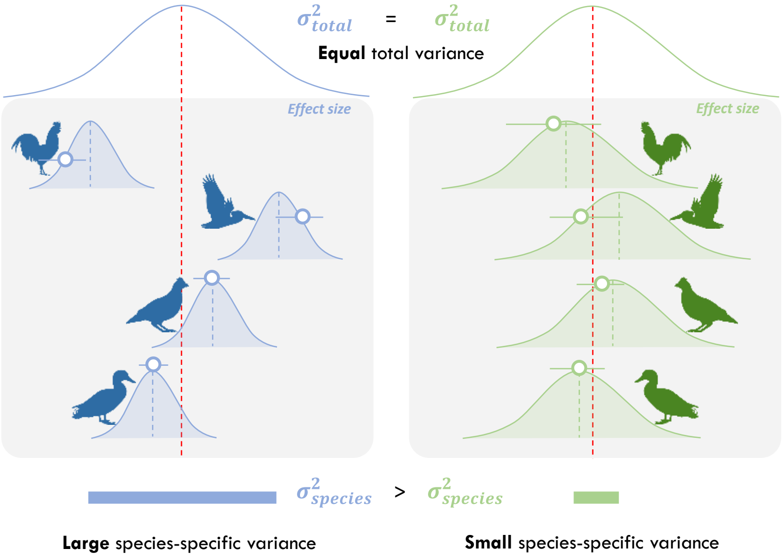
σ^2_{total} = σ^2_{total}
Equal total variance
Effect size
Effect size
$\sigma^2_{species}$ > $\sigma^2_{species}$
Large species-specific variance
Small species-specific variance

Fig. 2:

A cross-taxa meta-analysis with a high total variance can have a small amount of species level heterogeneity. The focal effect is still possible to be generalizable at the species level. The circles represent the replication of species-specific effects. The red dashed lines denote the meta-analytic mean effects. See a real example in **Extended strategies: Non-phylogenetic and phylogenetic species-level heterogeneity and generality**.

Here, we present a pluralistic framework designed to quantify heterogeneity, incorporating two intertwined strategies: stratification and the estimation of complementary measures of heterogeneity. We begin by introducing a general method for stratifying heterogeneity, which is applicable to any effect-size metric. We then evaluate commonly used heterogeneity metrics and propose two sets of new metrics, which capture different dimensions of heterogeneity and inform cross-context generalizability of the meta-analytic mean effect size. To ground our framework empirically, we undertake a large-scale synthesis, generating new benchmarks for interpreting heterogeneity and generalizability (Table 1), leveraging a big dataset spanning 512 ecological and evolutionary meta-analyses (cf. O'Dea *et al.* 2021; Costello & Fox 2022). We also present meta-scientific evidence on (in)congruence between different heterogeneity metrics, and outline approaches for developing useful extensions of heterogeneity quantification in phylogenetic contexts. To facilitate researchers in navigating the intricate landscape of heterogeneity, we conclude by offering practical recommendations and a tutorial with R functions (https://yefeng0920.github.io/heterogeneity_guide/). The proposed framework and large-scale synthesis aim to empower researchers in their quest to unravel the complex patterns underlying the generalizability of ecological and evolutionary phenomena.

Methods

Meta-analysis database

The ecological and evolutionary databases used in this study were originally compiled by Costello & Fox 2022, and O'Dea *et al.* 2021. They systematically searched for meta-analysis papers published in ecological journals, including those from the Ecological Society of America and journals of the British Ecological Society. Additionally, they supplemented the database with high-profile journals, such as Nature, and Science. Their systematic search yielded 522 meta-analysis datasets. We dropped meta-analysis datasets that could not achieve convergence when fitted to the multilevel model. Convergence could not be reached for ten meta-analysis datasets, even after adjusting key parameters of the iterative methods to maximize the log-likelihood function (see below for details). Therefore, our database contained 512 meta-analysis datasets encompassing 17,770 primary studies and 109,495 effect size estimates. On average, each meta-analysis dataset included 240 effect size estimates sourced from 40 studies, with median values of 64 and 23, respectively.

Stratifying heterogeneity using multilevel meta-analytic modelling framework

Data used in meta-analyses often exhibit a complex hierarchical structure (Nakagawa & Santos 2012; Noble *et al.* 2017), with paper (or study) identity serving as a typical clustering variable, forming two strata (or more). Ecological and evolutionary meta-analyses typically report around eight effect sizes per study (Yang *et al.* 2023). However, traditional random-effects meta-analytic approaches do not account for heterogeneity driven by such data stratification (Noble *et al.* 2022; Yang *et al.* 2022; Nakagawa *et al.* 2023), and multi-level meta-analysis is required to model heterogeneity at different strata or multi-levels in a meta-analysis (see **Appendix for the theoretical background**).

In the simplest multilevel model, the effect size estimate $ES_{[i]}$ is modelled as a combination of the population mean effect or meta-analytic mean effect size μ, random effects at two strata (i.e., between- and within-study levels), and statistical noise:

$$ES_{[i]} = \mu + u_{between[j]} + u_{within[i]} + e_{[i]}, (1)$$

The typical assumptions for Equation 1 are as follows: (i) between-study-level random effect $u_{b[j]}$ follows a normal distribution with mean zero and variance $\sigma^2_{between}$: $u_{between[j]} \sim \mathcal{N}(0, \sigma^2_{between})$, (ii) within-study-level random effect $u_{within[i]}$ follows a normal distribution with mean zero and variance σ^2_{within}: $u_{within[i]} \sim \mathcal{N}(0, \sigma^2_{within})$, and (iii) sampling error $e_{[i]}$ follows a normal distribution with mean zero and variance in effects defined by the sampling variance ($v_{[i]}$) associated with each effect size *i,* such that $e_{[i]} \sim \mathcal{N}\left(0, v_{[i]}\right)$. The assumption of homogeneous variances for the random effects can be relaxed to allow for heteroscedasticity (Viechtbauer & López‐López 2022). Similarly, the assumption of independent sampling errors ($e_{[i]}$) can be relaxed to allow for sampling error covariance $v_{[i]}$ (Noble *et al.* 2017; Yang *et al.* 2022). In the multilevel meta-analytic modelling framework, the total observed variance $\mathrm{Var}[ES_{[i]}]$ is the sum of the variance of true effects σ^2_{total} and the sampling variance, while the variance of true effects σ^2_{total} is the sum of between-study variance $\sigma^2_{between}$ and within-study variance σ^2_{within}. Note that in the context of random-effects model, the between-study variance (the so-called τ^2) is treated as the σ^2_{total}, while a multilevel model treats between-study variance as one of the components of the σ^2_{total}.

We used the *rma.mv()* function from the *metafor* package (Viechtbauer 2010) to fit all 512 meta-analysis datasets to the three-level meta-analytic model (Equation 1). We employed restricted maximum likelihood (REML) as the variance estimator and the quasi-Newton method as the optimizer to maximize the likelihood function over variance estimation

($\sigma^2_{between}$ and σ^2_{within}), with a threshold of 10^{-8}, a step length of 1, and a maximum iteration limit of 1000. All models successfully converged under these settings. We confirmed the identifiability of variance estimation ($\sigma^2_{between}$ and σ^2_{within}) by checking their likelihood profiles. The R code for model fitting can be accessed at the website (https://github.com/Yefeng0920/heterogeneity_ecoevo). In the following sections, we will elaborate on how to use Equation 1 to stratify heterogeneity information for different metrics.

Complementary measures of heterogeneity

Unstandardised heterogeneity metrics

Cochran's Q is a widely used metric for assessing heterogeneity in meta-analyses Cochran 1954. It serves as a test statistic to determine whether the true effects are homogeneous or not, informing a binary decision as to whether the effect sizes come from a common underlying population or not (i.e., is there variability around the true effect size?). In contrast, the variance of true effects ($\sigma^2_{total} = \sigma^2_{between} + \sigma^2_{within}$) provides a direct measure of absolute heterogeneity. Equation 1 offers a general way to partition the variance of the observed effects into sampling error variance, and that of true effects at different strata, such as between-study ($\sigma^2_{between}$) and within-study strata (σ^2_{within}). By considering additional strata, such as variation in effects among species or geographical locations, the total variance in true effects (σ^2_{total}) can be further decomposed to assess generalizability at these specific strata (See **Results and Discussion**). For example, low variation among species implies effects are similar, on average, across species. Nonetheless, relying solely on absolute variance may not provide practical intuition regarding the magnitude of effect heterogeneity. For example, in a meta-analysis with σ^2_{total} = 1, it is unclear whether this amount of variance is large and meaningful because absolute variance is not unitless and comparable across effect-size statistics.

Variance-standardised heterogeneity metrics

The heterogeneity index, I^2 has emerged as the most popular metric as it provides a standardized measure of heterogeneity that accounts for the scale dependence (i.e., unitless; Higgins *et al.* 2003). I^2 is a variance-scaled heterogeneity metric that measures the proportion of total variance beyond statistical noise (Higgins & Thompson 2002). The total I^2 can be computed by dividing the variance in the true effects (σ^2_{total}) by the variance in the observed effects ($\mathrm{Var}[ES_{[i]}]$), as follows:

$$I^2_{total} = \frac{\sigma^2_{total}}{\mathrm{Var}[ES_{[i]}]} = \frac{\sigma^2_{total}}{\sigma^2_{total} + \bar{v}}, (2)$$

where $\bar{v}$ represents the "typical" sampling error variance. $\bar{v}$ can be computed using different estimators (Takkouche, Cadarso-Suarez & Spiegelman 1999; Cheung 2014), with the common one being (Higgins & Thompson 2002):

$$\bar{v} = \frac{(k-1)\sum_{i=1}^{k} 1/v_{[i]}}{(\sum_{i=1}^{k} 1/v_{[i]})^2 - \sum_{i=1}^{k} 1/{v_{[i]}}^2}, (3)$$

Within the multilevel modelling framework, the total I^2 can be stratified at different strata (Nakagawa & Santos 2012; Cheung 2014), for example, by estimating I^2 at between-study ($I^2_{between}$) and within-study(I^2_{within}) levels:

$$I^2_{between} = \frac{\sigma^2_{between}}{\mathrm{Var}[ES_{[i]}]} = \frac{\sigma^2_{between}}{\sigma^2_{total} + \bar{v}}, (4)$$

$$I^2_{within} = \frac{\sigma^2_{within}}{\mathrm{Var}[ES_{[i]}]} = \frac{\sigma^2_{within}}{\sigma^2_{total} + \bar{v}}, (5)$$

However, as mentioned earlier, large I^2 values do not necessarily imply a practically relevant amount of heterogeneity (see Fig. 1; also see a case study in **Extended strategies: Non-phylogenetic and phylogenetic species-level heterogeneity and generality**). Statistical noise can sometimes inflate I^2 values, which is a common occurrence in ecology and evolutionary meta-analyses. Stratified I^2 metrics range from 0 to 100% (but together sum to 100%), providing a clearer intuition of the relative sources of heterogeneity and aiding in assessing the drivers of context dependence at different strata. For example, a I^2_{within} of 90% means within-study variation accounts for 90% of heterogeneity, therefore, indicating that within-study level predictors are more likely to drive context dependence. I^2 and its stratified variants can also be transformed into the ratio of the variance of true effect to typical sampling error variance ($\frac{\sigma^2}{\bar{v}} = \frac{I^2}{(1- I^2)}$ or $\log\left(\frac{\sigma^2}{\bar{v}}\right) = logit(\ I^2)$), which represents heterogeneity as a proportion of the statistical noise (sampling error variance).

Mean-standardised heterogeneity metrics

Evolutionary biologists and behavioural ecologists are familiar with the variance-scaled metrics such as heritability (h^2) and repeatability (R), which are statistically comparable to the heterogeneity index, I^2. Although less commonly used, there also exists the mean-scaled counterparts, such as evolvability or the coefficient of variation (CV) for additive genetic variance (CV_A) and CV for between-individual variance (CV_B) Hansen, Pélabon & Houle 2011. In a similar manner, there exists a mean-scaled heterogeneity metric that can provide a standardized measure of heterogeneity, denoted as CV_{total}, that compares the standard deviation σ_{total} to the magnitude of its mean population effect size (μ) (Takkouche, Cadarso-Suarez & Spiegelman 1999):

Commented [MNL1]: Total?

$$CV_{total} = \frac{\sigma_{total}}{|\mu|}, (6)$$

CV_{total} expresses the total heterogeneity as a proportion of the meta-analytic mean effect (or as a percentage of change in the meta-analytic mean effect when multiplied by 100). $CV_{total} = 1$ means that the heterogeneity (standard deviation among effects) is equal to mean population effect. Assuming a normal distribution this means ~16% of effects would have opposite sign to overall effect.

To provide a more precise quantification of heterogeneity at different strata, we propose stratified versions of CV_{total}. Under the simplest multilevel model framework (Equation 1), we propose estimating between-study, $CV_{between}$, and within-study, CV_{within}, as follows:

$$CV_{between} = \frac{\sigma_{between}}{|\mu|}, (7)$$

$$CV_{within} = \frac{\sigma_{within}}{|\mu|}, (8)$$

Notably, these mean-scaled variance metrics have the limitation of becoming arbitrarily large as the magnitude of meta-analytic mean effect $|\mu|$ approaches zero (Nakagawa *et al.* 2015). It is this limitation that has probably prevented the widespread adoption of the mean-scaled variance in the field of evolutionary quantitative genetic and animal personality research (Hansen, Pélabon & Houle 2011; Dochtermann & Royauté 2019).

Variance-mean-standardised heterogeneity metrics

To remedy the problems of I^2_{total} and CV_{total} as illustrated above, there is a more robust measure of heterogeneity M_{total} that combines the strengths of mean-scaled and variance-scaled metrics (Cairns & Prendergast 2022):

$$M_{total} = \frac{\sigma_{between} + \sigma_{within}}{\sigma_{between} + \sigma_{within} + |\mu|}, (9)$$

Here we propose between-study ($M_{between}$) and within-study (M_{within}) versions by stratifying M_{total}, which allows for a more precise quantification of heterogeneity at specific strata:

$$M_{between} = \frac{\sigma_{between}}{\sigma_{between} + \sigma_{witin} + |\mu|}, (10)$$

$$M_{within} = \frac{\sigma_{within}}{\sigma_{between} + \sigma_{within} + |\mu|}, (11)$$

M_{total} and its stratified variants are still standardised measures that quantify the size of heterogeneity relative to the magnitude of meta-analytic mean effect, providing intuitive interpretation. For example, $\sigma_{total} = 0$ leads to $M_{total} = 0$, indicating the population mean effect is fully generalisable, and replicable across different contexts (see a case study in **Extended strategies: Non-phylogenetic and phylogenetic species-level heterogeneity and generality**). One the other hand, M_{total} and its stratified variants are truncated at one, which overcomes the issue of CV_{total} when the magnitude of meta-analytic mean effect $|\mu|$ approaches zero. Note that there is another mean- and variance-scaled metric, M_{total}^2, where σ_{total} and $|\mu|$ are replaced by their squared values (See **Appendix**). CV_{total}, M_{total} and M_{total}^2 can be all be easily stratified using multilevel meta-analytic models.

Results and Discussion

Empirical patterns of heterogeneity and implications for effect generality

Source of heterogeneity

We used the variance-standardised metric I^2 to measure sources of heterogeneity. The 25th, 50th, and 75th percentiles corresponded to 79%, 91%, and 97% I^2_{total}, respectively (Fig. 3), which is worth contrasting with the conventional thresholds for interpreting I^2, which typically categorize heterogeneity as small, moderate, or high at 25%, 50%, and 75% I^2_{total} (Higgins *et al.* 2003), respectively. Thus, on average (50th percentile), 91% of variance in effect sizes can be attributed to the 'true' biological or methodological differences in research contexts, and may therefore be explainable using appropriate predictors. It also means that variation in true effect sizes is ten times larger than typical sampling error variance ($\frac{\sigma^2}{\bar{v}} = \frac{I^2}{(1- I^2)} = 10$; see Figs. S1 and S2 for empirical distributions of σ^2 and $\bar{v}$).

While I^2_{total} displayed a left-skewed and single-modal distribution, its stratified counterparts, $I^2_{between}$ and I^2_{within}, demonstrated a right-skewed distribution with multi-modal patterns (Fig. 3). There was no consistent trend suggesting neither type of stratified heterogeneity consistently outweighed the other across the 512 meta-analyses (Fig. 3). Intriguingly, 47% (242 out of 512) of the meta-analyses exhibited smaller between-study level heterogeneity than within-study level heterogeneity ($I^2_{between} < I^2_{within}$; Fig. 4). Within this subset of meta-analyses, the median values for I^2_{total}, $I^2_{between}$ and I^2_{within} were 95%, 21%, and 63%, respectively.

Our results highlight a key finding often overlooked by traditional heterogeneity quantification practices: findings from many meta-analyses with high total heterogeneity can still be generalized at the between-study study level. Such generalization is achievable when replication is defined as the testing of the null hypothesis at the between-study level, and when within-study methodological and biological variations can be adequately accounted for (i.e., within-lab heterogenization; Richter 2017) because some meta-analyses have relatively low heterogeneity at the between-study study level.

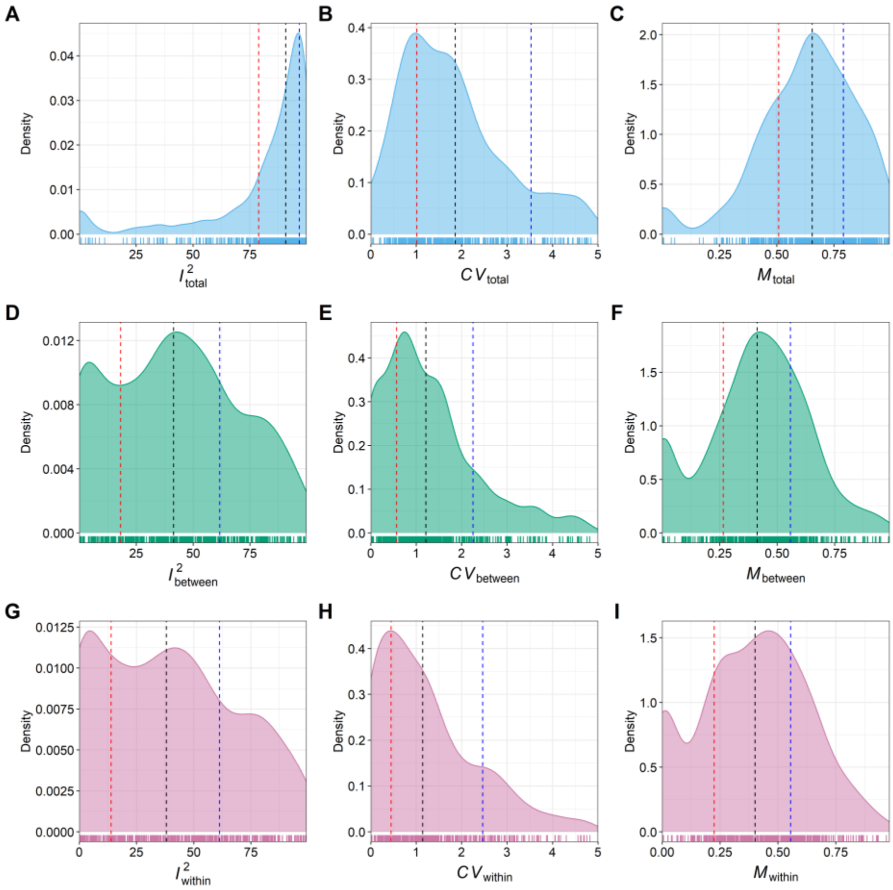

Fig. 3:

The distribution of heterogeneity estimates derived from 512 meta-analyses was systematically assessed using pluralistic measures and stratified across different strata. Total heterogeneity measures (A – C): I^2_{total}, CV_{total} and M_{total}. Between-study heterogeneity measures (D – E): $I^2_{between}$,

$CV_{between}$ and $M_{between}$. Within-study heterogeneity measures (G – I): I^2_{wihtin}, CV_{within} and M_{within}.Three dashed lines correspond to the 25th, 50th, and 75th percentiles, respectively. In panels B, E, and H, the CV was truncated at five for figure clarity, as very large CV values can be challenging to interpret when the meta-analytic mean effect is small. For example, the maximum CV observed in the 512 meta-analyses was 106, which was inflated by a small meta-analytic mean effect of 0.03. For the figures without truncation, please refer to Figure S3.

Magnitude of heterogeneity

When the mean-standardised metric CV_{total} was used to quantify the magnitude of heterogeneity, the calculated 25th, 50th, and 75th percentiles of CV_{total} values were 1.0, 1.8, and 3.5, respectively (Fig. 3). Therefore, the standard deviation (in this case, heterogeneity) was, on average (50-th percentile), nearly twice that of the meta-analytic mean effect. The distributions of both CV_{total} and its stratified versions, $CV_{between}$, and CV_{within}, displayed a right-skewed pattern with a single-mode (Fig. 3). In contrast, the distribution of the mean-variance-standardised metric M_t exhibited a more symmetrical pattern, with the 25th, 50th, and 75th percentiles of M_{total} values being 0.5, 0.6, and 0.8, respectively (Fig. 3), albeit with a minor peak around zero.

Notably, stratification analysis revealed that $M_{between}$ and M_{within} had patterns similar to those observed for $CV_{between}$ and CV_{within}. This similarity is expected as they can be mathematically transformed into one another using equations $M_{total} = CV_{total}/(1 + CV_{total})$ and $logit(M_{total}) = \log{(CV_{total})}$. The median values for both CV_{total} and M_{total} across the 512 meta-analyses signify a high amount of heterogeneity, thereby warranting a thorough exploration into the drivers influencing such context dependence. However, stratification of M_{total} also suggests that meta-analyses with high heterogeneity can possess a considerable likelihood of generality at the between-study level, given low $M_{between}$ (as we pointed out above with I^2). On average, there was a median $M_{between}$ = 0.3 (SD is 41% of the meta-analytic mean effect) observed in 47% of the meta-analyses (242/512) with smaller $M_{between}$ values compared to M_{within} values (Fig. 4).

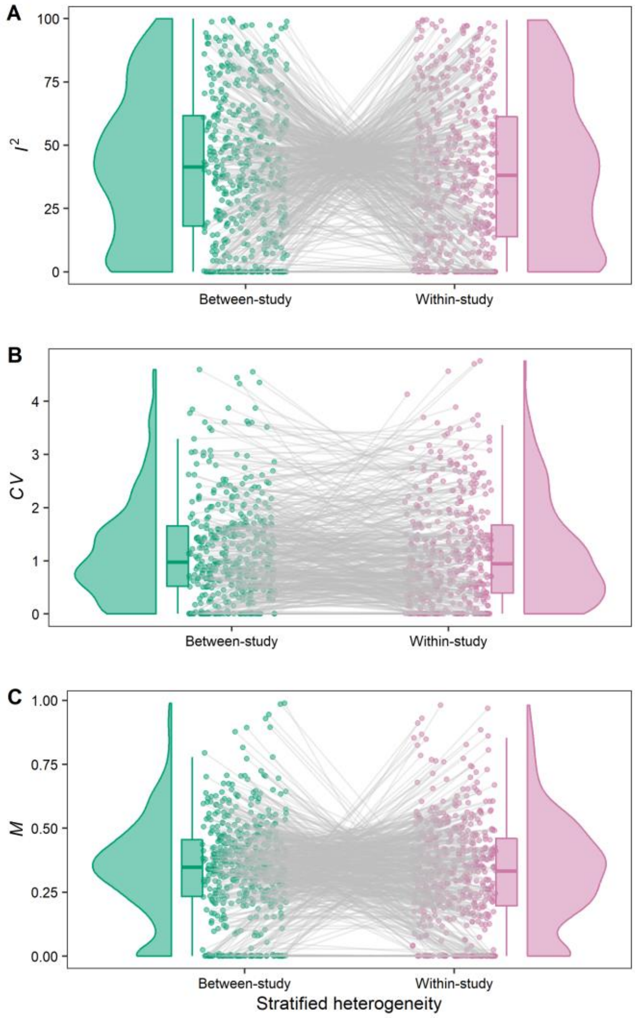

Fig. 4:

Paired comparison of stratified heterogeneity estimates derived from 512 meta-analyses for three heterogeneity metrics (A) I^2, (B) coefficient of variation, CV and (C) M. Heterogeneity was stratified at both ‘between-study’ and ‘within-study’ levels (x-axes). Each point represents an estimate from each meta-analysis. For panel B, CV has been truncated at five for figure clarity. For the full figures without truncation, please refer to Figure S4. For other details see Fig. 3.

Meta-scientific evidence on (in)congruence between different metrics

We found only moderate agreement between heterogeneity measured as I^2 and the alternatives (CV_{total}: r_{spearman} = 0.32, 95% CI = [0.24, 0.40], M_{total}: r_{spearman} = 0.33, 95% CI = [0.25, 0.41]; Fig. 5). In cases of meta-analyses with I^2 larger than 75% or smaller than 25% (identified as large and small heterogeneity by conventional benchmarks Higgins *et al.* 2003), the disagreement between I^2 and CV, as well as I^2 and M, became even more pronounced (Fig. S5 – S7). In contrast, a near-perfect agreement was observed between CV_{total} and M_{total}, as expected (r_{spearman} = 1, 95% CI = [0.99, 1]; Fig. 4). Therefore, cross-meta-analysis (meta-scientific) evidence suggests that I^2 as a measure of heterogeneity is not consistent with magnitude measures (CV_{total} and M_{total}) for ecological and evolutionary data. We also found that out of the 512 meta-analyses featuring medium to large I^2_{total} values (>50% based on conventional guidelines), 80 had small CV_{total} (Fig. 5), indicating that more than 20% of the large I^2_{total} values were caused by small sampling errors rather than larger amount of heterogeneity. These findings emphasize the importance of considering multiple metrics to obtain a holistic understanding of heterogeneity in meta-analyses (see **Interpreting heterogeneity and discerning effect generality using a pluralistic framework**).

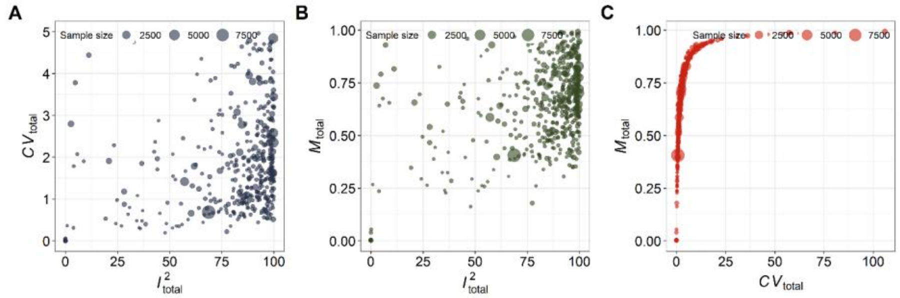

Fig. 5:

Disagreement (or agreement) between different heterogeneity metrics. For other details see Fig. 3. The Spearman correlation estimates (r_{spearman}) were: 0.32, 95% CI = [0.24, 0.40] for I^2_{total} and CV_{total}, 0.33, 95% CI = [0.25, 0.41] for I^2_{total} and M_{total}, and 1, 95% CI = [0.99, 1] forM_{total} and CV_{total}.

Extended strategies: Non-phylogenetic and phylogenetic species-level heterogeneity and generality

In ecological and evolutionary datasets, complexity often arises from the inclusion of diverse species, temporal, and spatial variations (Gurevitch *et al.* 2018). To address the challenge of quantifying heterogeneity in ecological and evolutionary datasets with increasingly complex structures that often involve high species-level heterogeneity, we propose decomposing heterogeneity into non-phylogenetic and phylogenetic species level strata. Such an approach offers a unique opportunity for further disentangling heterogeneity.

This can be achieved by embracing a flexible random-effects structure within the multilevel meta-analytic framework (Yang *et al.* 2022; Nakagawa *et al.* 2023). To illustrate this, we will show the principles of how to partition heterogeneity in datasets featuring multiple species (similar principles can be applied to those involving different temporal and spatial contexts). In the case of datasets encompassing multiple species, incorporating species-relevant random-effects terms into Equation 1 would lead to the phylogenetic multilevel meta-analytic model (Nakagawa & Santos 2012; Cinar, Nakagawa & Viechtbauer 2022):

$$ES_{[i]} = \mu + u_{species[k]} + u_{phylogeny[k]} + u_{between[j]} + u_{within[i]} + e_{[i]}, (12)$$

where $u_{species[k]}$ denotes the non-phylogenetic species random effect, which follows a normal distribution with mean zero and variance $\sigma^2_{species}$; $u_{phylogeny[k]}$ denotes the phylogenetic species random effect, which follows a multivariate normal distribution with mean zero and variance-covariance matrix $\sigma^2_{phylogeny}\boldsymbol{A}$ (where $\sigma^2_{phylogeny}$ is the phylogenetic species variance, and $\boldsymbol{A}$ is phylogenetic correlation matrix based on the distance between species on a molecular-based phylogenetic tree).

With Equation 12 in hand, the total variance can be stratified at the phylogenetic and non-phylogenetic species level ($\sigma^2_{phylogeny}$ and $\sigma^2_{species}$). Such stratification allows for the assessment of the generality of a focal effect within these strata, as illustrated in the empirical example below. Phylogenetic and non-phylogenetic species-level heterogeneity can be measured using $I^2_{phylogeny}$ and $I^2_{species}$, respectively:

$$I^2_{phylogeny} = \frac{\sigma^2_{phypogeny}}{\sigma^2_{phypogeny} + \sigma^2_{species} + \sigma^2_{between} + \sigma^2_{within} + \bar{v}}, (13)$$

$$I^2_{species} = \frac{\sigma^2_{species}}{\sigma^2_{phylogeny} + \sigma^2_{sspecies} + \sigma^2_{between} + \sigma^2_{within} + \bar{v}}, (14)$$

We derive the alternative stratified version of measures as follows:

$$CV_{phylogeny} = \frac{\sigma_{phylogeny}}{|\mu|}, (15)$$

$$CV_{species} = \frac{\sigma_{species}}{|\mu|}, (16)$$

$$M_{phylogeny} = \frac{\sigma_{phylogeny}}{\sigma_{phylogeny} + \sigma_{species} + \sigma_{between} + \sigma_{within} + |\mu|}, (17)$$

$$M_{species} = \frac{\sigma_{species}}{\sigma_{phylogeny} + \sigma_{species} + \sigma_{between} + \sigma_{within} + |\mu|}, (18)$$

To illustrate the insights gained through these extended measures, we present an empirical example. We re-analysed a phylogenetic meta-analysis originally conducted by (Risely et al. Risely, Klaassen & Hoye 2018). Our focus centres on a subset of this analysis, specifically examining the impact of infection status on the cost (e.g., movement capacity) of migratory animals. Our re-analysis yielded three observations. Firstly, I^2_{total}= 97% exceeded the 75th percentile of the empirically derived heterogeneity distribution (Fig. 6 and Table 1). This suggests a high amount of heterogeneity according to the conventional benchmarks (Higgins *et al.* 2003). However, when we employed magnitude metrics to measure heterogeneity, they fell below the 25th and 50th percentiles of the empirically derived heterogeneity distribution (CV_{total}= 1.3 and M_{total}= 0.6). This discrepancy was attributed to the small typical sampling variance $\bar{v}$, which was found to be 0.001 in this case, underscoring I^2_{total}'s limitation of relying on $\bar{v}$ to capture relative magnitude of heterogeneity. On the other hand, we emphasise that the proper interpretation of I^2_{total} is to use it to indicate the source of heterogeneity rather than the magnitude, as it represents the variance of the true effect in the context of the variance of the observed effect. For example, I^2_{total}= 97% suggests a heterogeneity can explain most (97%) of the variability in effect size (only 3% is explained by the sampling variance, or the heterogeneity is 32 times larger than that of statistical noise).

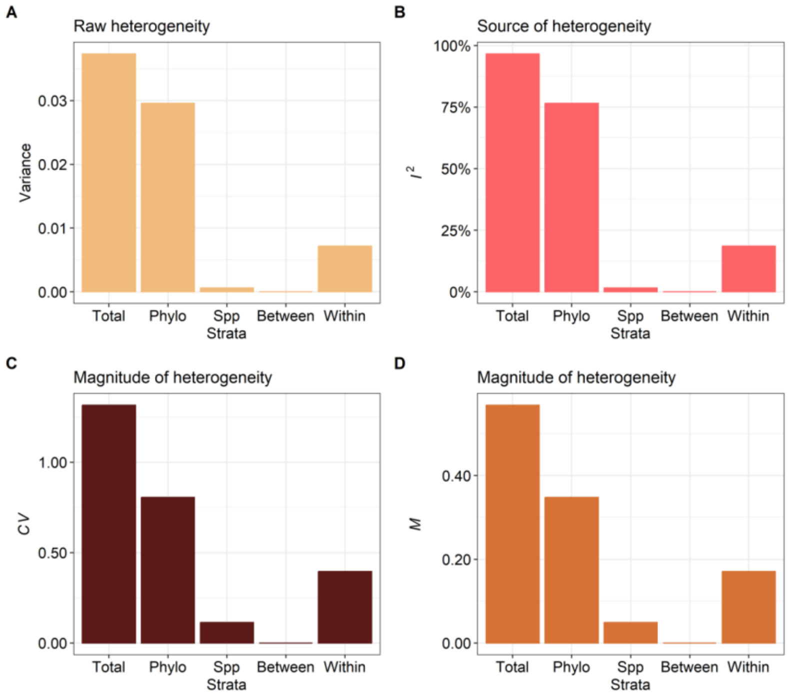

Fig. 6:

Heterogeneity quantification and stratification for multiple metrics. (A) The heterogeneity is quantified using raw variance, (B) source measure I^2, (C) magnitude measure CV, and (D) magnitude measure M, and stratified at phylogenetic (Phylo), non-phylogenetic (Spp), between-study (Between), and within-study (Within) levels. The source measure I^2 sometimes aligns well with the raw variance, as observed in this example (A and B). However, we note that I^2 values can be challenging to interpret as the magnitude of heterogeneity, especially when the typical sampling error variance is extremely small or large. This challenge is often encountered with certain effect size measures, such as the log coefficient of variation ratio (lnCVR), as demonstrated in a real example at https://yefeng0920.github.io/heterogeneity_guide/.

Secondly, the estimated mean effect was highly likely to be generalizable and replicable at the between-study- and species-context, if controlling for within-study experimental contexts (e.g., age, sex, outcomes). This is indicated by the stratification analysis that between-study level heterogeneity was extremely low, despite a large heterogeneity according to conventional benchmarks (Higgins *et al.* 2003). Traditional meta-analytic practices would overlook these valuable insights, potentially leading to erroneous conclusions. For example, random-effects meta-analysis shows that this dataset has high study-level heterogeneity (I^2_{total} = 96%; Fig. 4 and Table 1). However, stratification of heterogeneity further indicated that it was not attributable to the between-study level but, rather, was mainly explained by the phylogenetic signal ($I^2_{phylogeny}$ = 76%).

Interpreting heterogeneity and discerning effect generality using a pluralistic framework

Given that two strategies for heterogeneity quantification (i.e., new metrics and stratification of heterogeneity) offer distinct insights into empirical patterns of biological generality (**Figs. 2** to **7**), we propose adopting a pluralistic framework to comprehensively assess generality by more thoroughly characterising and presenting meta-analytic heterogeneity. Our recommendations are fourfold (**Table 1**):

(1) Employ a multilevel meta-analytic framework: We strongly advocate for the use of a multilevel meta-analytic framework (Equation 1), as opposed to random-effects meta-analysis, for the modelling and stratification of heterogeneity. Additional random effects can be incorporated into Equation 1 as needed to further dissect heterogeneity. For example, the application of the phylogenetic multilevel meta-analytic model (Equation 12) allows for the disentanglement of species-specific heterogeneity.

(2) Quantification and stratification of pluralistic heterogeneity measures: We recommend transparently reporting all variance components, including typical sampling error variances in the main text, supplementary tables, or figures (**Figs. 6** and **7** and **Table 1**). As such, pluralistic metrics can be computed using the formula above. I^2, M (with CV being derivable from M), and their stratified versions should be reported as the default measures. These measures provide complementary information, for example, the source and magnitude (examples see **Table 1**). We also provide parametric bootstrapping solutions to estimate the uncertainty (e.g., 95%CI) for each of the measures.

(3) Check the model parameter identifiability: When models incorporate many random effects, issues of parameter identifiability may arise, wherein unique variance estimates that maximize the likelihood function may not exist (see **Methods**; Raue *et*

al. 2009). Therefore, we recommend assessing whether variance components are all identifiable through means such as checking profile likelihood, before proceeding with heterogeneity quantification and stratification.

(4) Carefully interpret heterogeneity measures: It is important to interpret both total and stratified heterogeneity to evaluate variation in effect sizes, aiding in the examination of general rules in the fields of ecology and evolution. However, neither the conventional benchmarks (25, 50, and 75% as small, moderate and high heterogeneity; Higgins *et al.* 2003) nor those of empirically derived distributions (**Table 1** and **Fig. 3**) are currently suitable for informing interpretation. Nevertheless, the empirically derived distribution can be employed to interpret heterogeneity within the context of existing ecological and evolutionary meta-analyses.

Overall, we argue that ecologists and evolutionary biologists should treat heterogeneity and the meta-analytic mean effect size with equal importance and discuss both when making biological conclusions (Higgins, Thompson & Spiegelhalter 2009). Our pluralistic approach provides a framework to achieve it.

Table 1

Summary of heterogeneity measures, their stratified counterparts, and empirically derived benchmark values. SMD denotes standardised mean difference. lnRR denotes log response ratio. *Zr* denotes Fisher's r-to-z transformed correlation coefficient. 2-by-2 table denotes often dichotomous (binary) effect size measures, such as log odds ratio, log risk ratio. Uncommon measures represent less frequently used effect size measures, such as raw mean difference and regression coefficients.

Types	Metrics	Interpretation and examples	Empirically derived benchmark[1]
Test statistic	Q	Null-hypothesis test. Statistical test of heterogeneity in effect sizes.	Not applicable
Unstandardisation	σ^2	Absolute magnitude measure of heterogeneity. Variance (square of standard deviation) of the meta-analytic mean effect (σ^2_{total}) and its stratification at between- and within-study contexts ($\sigma^2_{between}$ and σ^2_{within}).	25th, 50th, and 75th percentiles (Fig. S1): 0.54, 1.25, 3.03 for SMD; 0.11, 0.27, 0.57 for lnRR; 0.06, 0.12, 0.25 for *Zr*; 1.04, 1.20, 2.51 for the 2-by-2 table; 0.01, 0.04, 0.27 for uncommon measures. The percentiles of typical sampling variance $\bar{v}$ are reported at Fig. S2.
Variance-standardization	I^2	Heterogeneity source measure. Proportion of variance not due to statistical noise. It measures the source of heterogeneity. For example, σ^2_{total} = 95% denotes that 95% of variation is the result of nuisance heterogeneity (i.e., differences in contexts). $\sigma^2_{between}$ = 80% and σ^2_{within} = 15% indicate differences in between-study contexts dominate the heterogeneity, pointing towards between-study level predictors as the likely drivers of context-dependent variation.	25th, 50th, and 75th percentiles (Fig. 3): 79%, 91%, 97% for overall; 78%, 89%, 96% for SMD; 88%, 95%, 99% for lnRR; 73%, 87%, 95% for *Zr*; 71%, 73%, 89% for the 2-by-2 table; 74%, 91%, 98% for uncommon measures.
Mean-standardization	CV	Heterogeneity magnitude measure. Variance expressed as the proportion of the mean effect. It is the measure of the magnitude of heterogeneity in	25th, 50th, and 75th percentiles (Fig. 3):

		the context of mean effect. For example, CV_{total} = 1.5, $CV_{between}$ = 0.8, and CV_{within}= 0.5 denote that total, between- and within-study variance are 150, 80, and 50% of the mean effect.	1.0, 1.8, 3.5 for overall; 1.1, 2.0, 3.9 for SMD; 1.2, 1.9, 3.5 for lnRR; 0.8, 1.7, 2.9 for *Zr*; 1.2, 2.2, 2.7 for the 2-by-2 table; 0.7, 1.1, 1.3for uncommon measures.
Variance-mean-standardization	M	Heterogeneity magnitude measure. Variance expressed as the proportion of the mean effect and a transformation of CV designed with better properties. It is the measure of the magnitude of heterogeneity in the context of mean effect. The interpretation can be eased by back-transformation with $M_{total} = CV_{total}/(1 + CV_{total})$. For example, CV_{total} = 0.6, $CV_{between}$ = 0.5, and CV_{within}= 0.4 denote that total, between- and within-study variance are 150, 100, and 67% of the mean effect.	25th, 50th, and 75th percentiles (Fig. 3): 0.5, 0.7, 0.8 for overall; 0.5, 0.7, 0.8 for SMD; 0.5, 0.7, 0.8 for lnRR; 0.5, 0.6, 0.8 for *Zr*; 0.5, 0.7, 0.7 for the 2-by-2 table; 0.4, 0.5, 0.6 for uncommon measures.

[1] The distributions and percentiles could be underestimated if publication bias existed.

Appendix

Stratifying heterogeneity of hierarchical meta-analytic data

In this section, we elucidate the theoretical background behind employing a three-level meta-analytic approach to stratify datasets characterized by three-level hierarchical structure as outlined above. Note that the stratification of heterogeneity can be further extended to data structures with more than four strata as necessary (see a case study in **Extended strategies: Non-phylogenetic and phylogenetic species-level heterogeneity and generality**). In the first-stage modelling procedure, the true (population) effect size $\mu_{between[j]}$ of j-th study is modelled using a normal distribution with expectation μ and variance $\sigma^2_{between}$, where μ is the population mean effect or overall effect and $\sigma^2_{between}$ denotes the extent to which $\mu_{between[j]}$ deviates from the overall effect μ Van den Noortgate *et al.* 2013; Cheung 2014. Moving to the second-stage modelling procedure, the i-th effect size $\mu_{within[i]}$ within j-th study is modelling using a normal distribution with expectation $\mu_{between[j]}$ and variance σ^2_{within}, where σ^2_{within} represents the extent to which within-study effect $\mu_{within[i]}$ deviates from between-study effect $\mu_{between[j]}$Van den Noortgate *et al.* 2013; Cheung 2014. In the third-stage modelling procedure, the effect size estimate $ES_{[i]}$ of $\mu_{within[i]}$ is modelled using a normal distribution with expectation $\mu_{within[i]}$ and sampling error variance$v_{[i]}$. This multilevel modelling framework provides a general way to decompose the variance of effect sizes into different strata, for example between- and within-study levels.

From the implementation perspective, effect size estimate $ES_{[i]}$ is not sequentially modelled through the three-stage process but rather directly modelled from the overarching distribution with an expectation μ and variance-covariance matrix VCV Van den Noortgate *et al.* 2013; Cheung 2014:

$$\begin{bmatrix} \sigma^2_{between} + \sigma^2_{within} + v_{[1]} & \cdots & \sigma^2_{between} \\ \vdots & \ddots & \vdots \\ \sigma^2_{between} & \cdots & \sigma^2_{between} + \sigma^2_{within} + v_{[k]} \end{bmatrix}, (19)$$

The meta-analytic model specified with the variance-covariance matrix VCV is referred to as the multilevel meta-analytic model (Equation 1). VCV can be reparametrized as a compound symmetry random-effects structure within the framework of multivariate meta-analytic model Van den Noortgate *et al.* 2013; Cheung 2019.

$$\begin{bmatrix} \sigma^2_{total} + v_{[1]} & \cdots & \rho\sigma^2_{total} \\ \vdots & \ddots & \vdots \\ \rho\sigma^2_{total} & \cdots & \sigma^2_{total} + v_{[k]} \end{bmatrix}, (20)$$

where $\sigma^2_{total} = \sigma^2_{between} + \sigma^2_{within}$ is the total variance in effect sizes and $\rho = \sigma^2_{between}/\sigma^2_{total}$ denotes intraclass correlation coefficient.

Extended heterogeneity metrics

In addition to CV_{total}, M_{total}, and their stratified counterparts (Equations 6 – 11), we introduce two related heterogeneity measures. CV_{total} has a potential shortcoming that it is not numerically equivalent to the sum of heterogeneity at between- and within-study levels ($CV_{total} \neq CV_{between} + CV_{within}$). This is because the total standard deviation σ_t is not equal to the sum deviations at each stratum ($\sigma_{total} \neq \sigma_{between} + \sigma_{within}$). To address the numerical difference, we propose CV^2_{total}, an analogue to CV_{total}:

$$CV^2_{total} = \frac{\sigma^2_{total}}{\mu^2}, (21)$$

Similarly, we propose between-study level and within-study level variants ($CV^2_{between}$ and CV^2_{within}):

$$CV^2_{between} = \frac{\sigma^2_{between}}{\mu^2}, (22)$$

$$CV^2_{wihtin} = \frac{\sigma^2_{within}}{\mu^2}, (23)$$

Following the same principle, M_{total}^2 can be obtained Cairns & Prendergast 2022:

$$M_{total}^2 = \frac{\sigma_{total}^2}{\sigma_{total}^2 + \mu^2}, (24)$$

We further propose between-study level (M_{total}^2) and within-study level (M_{total}^2) counterparts as:

$$M_{between}^2 = \frac{\sigma_{between}^2}{\sigma_{total}^2 + \mu^2}, (25)$$

$$M_{within}^2 = \frac{\sigma_{within}^2}{\sigma_{total}^2 + \mu^2}, (26)$$

M_{total}^2 and its stratified variants ($M_{between}^2$ and M_{within}^2) are re-scaling of CV_{total}^2 and its stratified variants ($CV_{between}^2$ and CV_{within}^2). Therefore, they can be converted into each other using simple mathematical relationships, such as ${M_{total}^2}^{-1} = {CV_{total}^2}^{-1} + 1$ or $\text{logit}(M_{total}^2) = \log{(CV_{total}^2)}$.

References

Borenstein, M., Higgins, J.P., Hedges, L.V. & Rothstein, H.R. (2017) Basics of meta‐analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods,* **8,** 5-18.

Cairns, M. & Prendergast, L.A. (2022) On ratio measures of heterogeneity for meta‐analyses. *Research Synthesis Methods,* **13,** 28-47.

Cheung, M.W.-L. (2014) Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological methods,* **19,** 211.

Cheung, M.W.-L. (2019) A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology review,* **29,** 387-396.

Cinar, O., Nakagawa, S. & Viechtbauer, W. (2022) Phylogenetic multilevel meta‐analysis: A simulation study on the importance of modelling the phylogeny. *Methods in Ecology and Evolution,* **13,** 383-395.

Cochran, W.G. (1954) The combination of estimates from different experiments. *Biometrics,* **10,** 101-129.

Costello, L. & Fox, J.W. (2022) Decline effects are rare in ecology. *Ecology,* **103,** e3680.

Dochtermann, N.A. & Royauté, R. (2019) The mean matters: going beyond repeatability to interpret behavioural variation. *Animal Behaviour,* **153,** 147-150.

Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. (2018) Meta-analysis and the science of research synthesis. *Nature,* **555,** 175-182.

Hansen, T.F., Pélabon, C. & Houle, D. (2011) Heritability is not evolvability. *Evolutionary Biology,* **38,** 258-277.

Higgins, J.P. & Thompson, S.G. (2002) Quantifying heterogeneity in a meta‐analysis. *Statistics in medicine,* **21,** 1539-1558.

Higgins, J.P., Thompson, S.G., Deeks, J.J. & Altman, D.G. (2003) Measuring inconsistency in meta-analyses. *bmj,* **327,** 557-560.

Higgins, J.P., Thompson, S.G. & Spiegelhalter, D.J. (2009) A re‐evaluation of random‐effects meta‐analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* **172,** 137-159.

IntHout, J., Ioannidis, J.P., Rovers, M.M. & Goeman, J.J. (2016) Plea for routinely presenting prediction intervals in meta-analysis. *BMJ open,* **6,** e010247.

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

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

Nakagawa, S., Yang, Y., Macartney, E.L., Spake, R. & Lagisz, M. (2023) Quantitative evidence synthesis: a practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environmental Evidence,* **12,** 8.

Noble, D.W., Lagisz, M., O'dea, R.E. & Nakagawa, S. (2017) Nonindependence and sensitivity analyses in ecological and evolutionary meta‐analyses. *Molecular Ecology,* **26,** 2410-2425.

Noble, D.W., Pottier, P., Lagisz, M., Burke, S., Drobniak, S.M., O'Dea, R.E. & Nakagawa, S. (2022) Meta-analytic approaches and effect sizes to account for 'nuisance heterogeneity'in comparative physiology. *Journal of Experimental Biology,* **225,** jeb243225.

O'Dea, R.E., Lagisz, M., Jennions, M.D., Koricheva, J., Noble, D.W., Parker, T.H., Gurevitch, J., Page, M.J., Stewart, G. & Moher, D. (2021) Preferred reporting items for systematic reviews and meta‐analyses in ecology and evolutionary biology: a PRISMA extension. *Biological reviews,* **96,** 1695-1722.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U. & Timmer, J. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics,* **25,** 1923-1929.

Richter, S.H. (2017) Systematic heterogenization for better reproducibility in animal experimentation. *Lab animal,* **46,** 343-349.

Risely, A., Klaassen, M. & Hoye, B.J. (2018) Migratory animals feel the cost of getting sick: A meta‐analysis across species. *Journal of Animal Ecology,* **87,** 301-314.

Rücker, G., Schwarzer, G., Carpenter, J.R. & Schumacher, M. (2008) Undue reliance on I2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology,* **8,** 1-9.

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

Spake, R., O'dea, R.E., Nakagawa, S., Doncaster, C.P., Ryo, M., Callaghan, C.T. & Bullock, J.M. (2022) Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology & Evolution***,** 1-11.

Takkouche, B., Cadarso-Suarez, C. & Spiegelman, D. (1999) Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American journal of epidemiology,* **150,** 206-215.

Van den Noortgate, W., López-López, J.A., Marín-Martínez, F. & Sánchez-Meca, J. (2013) Three-level meta-analysis of dependent effect sizes. *Behavior research methods,* **45,** 576-594.

Viechtbauer, W. (2010) Conducting meta-analyses in R with the metafor package. *Journal of statistical software,* **36,** 1-48.

Viechtbauer, W. & López‐López, J.A. (2022) Location‐scale models for meta‐analysis. *Research Synthesis Methods,* **13,** 697-715.

Yang, Y., Lagisz, M., Williams, C., Pan, J., Noble, D.W. & Nakagawa, S. (2023) Robust point and variance estimation for ecological and evolutionary meta-analyses with selective reporting and dependent effect sizes. *EcoEvoRxiv*.

Yang, Y., Macleod, M., Pan, J., Lagisz, M. & Nakagawa, S. (2022) Advanced methods and implementations for the meta-analyses of animal models: Current practices and future recommendations. *Neuroscience & Biobehavioral Reviews***,** 105016.

Supplementary Materials

Table S1, Fig. S1 to Fig. S7.