Measuring biological generality in meta-analysis: a pluralistic approach to heterogeneity quantification and stratification

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

[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)

# Equal senior author

ORCID

Yefeng Yang 0000-0002-8610-4016

Daniel W. A. Noble 0000-0001-9460-8743

Rebecca Spake 0000-0003-4671-2225

Alistair M. Senior 0000-0001-9805-7280

Malgorzata Lagisz 0000-0002-3993-6127

Shinichi Nakagawa 0000-0002-7765-5182

Abstract

Uncovering general rules enhances the predictive capabilities in ecology and evolution. Meta-analytic approaches play a critical role in this endeavour, examining the extent to which phenomena can be replicated, generalized, and transferred. However, ecologists and evolutionary biologists have largely overlooked the role of meta-analytic heterogeneity in informing generality. To reform this situation, we introduce a pluralistic approach aimed at quantifying and stratifying various heterogeneity metrics, such as I^2, CV, M, and predictive distribution. These metrics offer complementary information, revealing the source, magnitude, and visual representation of heterogeneity. Our analysis of 512 meta-analyses demonstrates that heterogeneity is, on average, ten times larger than statistical noise, contributing to 91% of the observed variance (median I^2 = 91%). This amount of heterogeneity is nearly twice the size of the meta-analytic mean effect (median CV = 1.8, M = 0.6), indicating substantial total heterogeneity in ecology and evolution. Surprisingly, in half of the cases, focal effects could generalize across studies even with high total heterogeneity by controlling for within-study variation. Our synthesis also visualises empirical distributions of various heterogeneity metrics, potentially serving as new benchmarks for informed interpretation. Our proposed pluralistic approach will accelerate the future quest for general rules via meta-analyses.

Main

Uncovering general patterns holds immense significance in ecology and evolution [1]. This enables scientists, practitioners, and policymakers to transfer findings across diverse systems, taxonomic groups, and spatiotemporal contexts. This pursuit enhances predictive capabilities and facilitates more precise management, intervention, and conservation practices. Ecologists and evolutionary biologists strive to unveil general processes and patterns using a range of approaches [2]. Notably, meta-analytic modelling has emerged as a natural route to assess the generality or context dependence of an effect of interest. By synthesizing a collection of conceptual replications [3], meta-analyses can scrutinize the extent to which inferences drawn from a specific context can be replicated (replication), extended beyond the reference context to a new context of interest (transferred), and extrapolated to the broader target population (generalized) as requested by stakeholders [2,4].

Meta-analyses play a crucial role in evaluating the generality of patterns [3]. Firstly, they quantitatively estimate the population mean effect across studies [5-7], characterising the central tendency of a focal effect. Secondly, they can identify effect modifiers or moderators contributing to context dependence [5] and provide tailored estimates for target contexts [4]. Third, meta-analyses can quantify variability in study outcome, the “heterogeneity” among effect sizes. Without quantifying heterogeneity, it is difficult to interpret both the overall trends and context-specific effects [8]. Heterogeneity can help to indicate the degree of inconsistency, or context dependence, of study findings, with high heterogeneity signalling a need to investigate the drivers of the variation. Lower heterogeneity can indicate high generality. Specifically, the mean effect size is highly transferable across the contexts characterised by the study pool without the need to consider effect modifiers [2]. Until now, the significance of heterogeneity in informing generality has been largely overlooked. Indeed, surveys have revealed that heterogeneity statistics are not routinely reported [7-9].

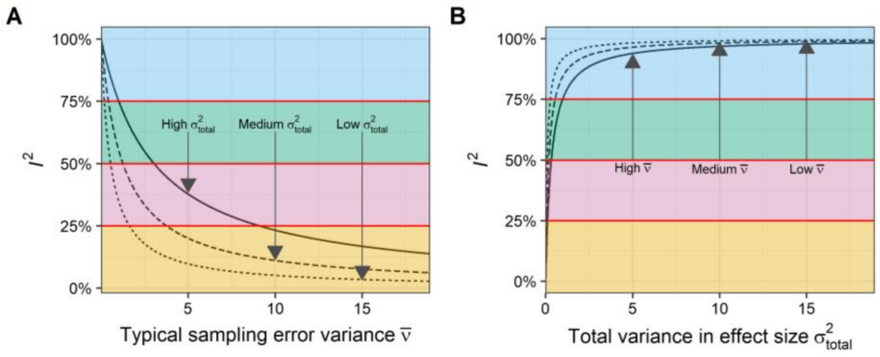
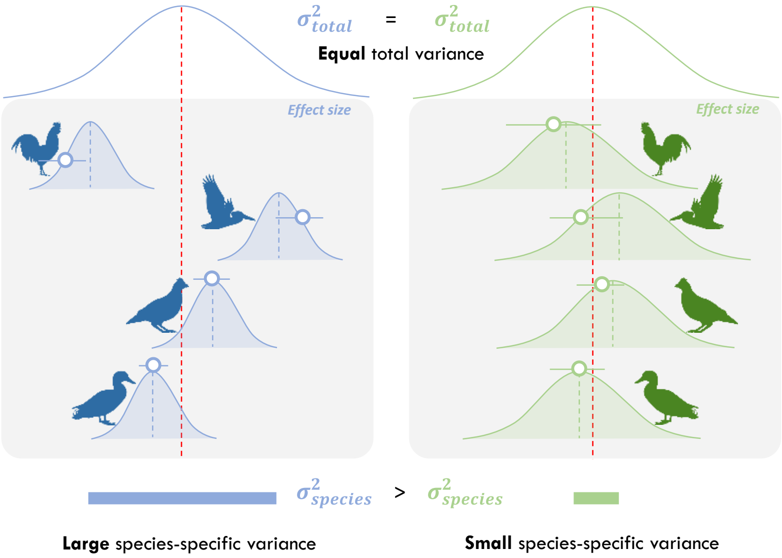


Fig. 1:

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 sampling error variances $\bar{v}$ (i.e., low statistical noise). (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 [10].

Currently, measuring and interpreting meta-analytic heterogeneity faces two major limitations. First, no single heterogeneity metric provides a holistic interpretation of generality [11].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) [12,13]. The biological interpretation of I^2, however, is ambiguous [14] because a small absolute heterogeneity can lead to a high I^2 due to small statistical noise (see Fig. 1) [12,14,15]. In addition, I^2 is a point estimate and cannot reflect the whole distribution of context-specific effects [16]. Second, meta-analyses typically focus on estimating total heterogeneity only [5], despite the hierarchical nature of real biological data structures [6,9]. Explicitly decomposing effect size heterogeneity across hierarchical levels (i.e., stratification) enables a more nuanced assessment of generality, and helps in identifying contextual factors [5] 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 [8].



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-specific heterogeneity. The focal effect is still possible to be generalizable at the species level. The circles represent the replication of species-specific effect. The red dashed lines denote the meta-analytic mean effects. See a real example in **Modelling additional source heterogeneity**.

Here, we present solutions to the aforementioned limitations, offering pluralistic pathways to biological generality and transferability. We begin by reformulating the concept of heterogeneity within the multilevel meta-analytic model and evaluating commonly used heterogeneity measures. Building on this foundation, we take currently underused heterogeneity metrics and propose new, stratified versions. After introducing the theoretical background, we leverage a big dataset spanning 512 meta-analyses from the fields of ecology and evolutionary biology (cf. [17,18]) to unveil empirical patterns of heterogeneity using these measures and establish meta-scientific evidence on their (in)congruence. Next, we show ways to visualise measures of heterogeneity using predictive distributions. Finally, we provide practical recommendations and a tutorial with R functions for researchers to navigate the complex landscape of heterogeneity (https://yefeng0920.github.io/heterogeneity_guide/). Our synthesis highlights the significance of adopting a pluralistic framework for a comprehensive understanding of meta-analytic findings in ecology and evolutionary biology.

Discerning biological generality

Heterogeneity in multilevel meta-analytic modelling framework

Data used in meta-analyses often exhibit a complex hierarchical structure [5,19], with 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 [20]. However, Traditional random-effects meta-analytic approaches do not account for heterogeneity driven by such data stratification [6,7,9], and multi-level meta-analysis is required to model heterogeneity at different strata or multi-levels in a meta-analysis (see **Methods**).

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 is 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 [21]. Similarly, the assumption of independent sampling errors ($e_{[i]}$) can be relaxed to allow for sampling error covariance $v_{[i]}$ [7]. In the following sections, we will elaborate on how to stratify heterogeneity information using Equation 1.

Unstandardised heterogeneity metrics

Cochran's Q is a widely used metric for assessing heterogeneity in meta-analyses [22]. 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 heterogeneity 'non-zero'?). 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 generality at these specific strata (See **Model additional source heterogeneity**). For example, high variation among studies implies lack of generality from one study to another while low variation among species implies effects are similar, on average, across species. Nonetheless, relying solely on such absolute variance may not provide practical intuition regarding the magnitude of 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 unit-free and not comparable across effect size measure used.

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., unit-free) [10]. I^2 is a variance-scaled heterogeneity metric that measures the proportion of total variance beyond statistical noise [13]. 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 [23,24], with the common one being [13]:

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

Within the multilevel modelling framework, the total I^2 can be stratified at different strata [5,24], for example, by estimating I^2 at between-study ($I^2_{between}$) and within-study(I^2_{within}) levels:

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

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

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 **Model additional source of heterogeneity**). Statistical noise can sometimes inflate I^2 values, which is a common occurrence in ecology and evolutionary meta-analyses (see **Empirical patterns of heterogeneity in ecology and evolution**). Stratified I^2 metrics range from 0 to 100% (but together sum to 100%), providing a clearer intuition of the source 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 can account 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) [25]. 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_t to the magnitude of its population mean (μ) [23]:

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

CV_t 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). To provide a more precise quantification of heterogeneity at different strata, we propose stratified versions of CV_t. Under the simplest multilevel model framework (Equation 1), we propose estimating between-study, CV_b, and within-study, CV_w, 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 [26]. 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 [25,27].

Variance-mean-standardised heterogeneity metrics

To remedy the problems of I^2 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 [11]:

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

Here we propose between-study ($M_{between}$) and within-study (M_{within}) versions by stratifying M_t, 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_t 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 **Model additional source of heterogeneity**). 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^2_{total}, where σ_{total} and $|\mu|$ are replaced by their squared values (**Methods**). CV_{total}, M_{total} and M^2_{total} can be all be easily stratified using multilevel meta-analytic models (**Model additional source of heterogeneity**).

Empirical patterns of heterogeneity in ecology and evolution

To evaluate empirical patterns in heterogeneity among meta-analytic studies in ecology and evolution, we applied multilevel meta-analytic models (Equation 1) to 512 published meta-analyses [18,28]. For each meta-analysis, we quantified and stratified heterogeneity using I^2_{total}, CV_{total}, M_{total}. For I^2_{total}, the 25th, 50th, and 75th percentiles corresponded to 79%, 91%, and 97% I^2_{total}, respectively (Fig. 3), rather than conventional thresholds for interpreting I^2, which typically categorize heterogeneity as small, moderate, or high at 25%, 50%, and 75% I^2_{total}, respectively [10]. This also means, on average, variation in true effect sizes σ^2 was ten times as large as 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 $\bar{v}$) and 91% of them can be attributed to the 'true' biological or methodological differences in research contexts, and thus are theoretically explainable using appropriate predictors.

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. There was no consistent trend suggesting one 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. It highlights 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 [29]) because some meta-analyses have relatively low heterogeneity at the between-study study level.

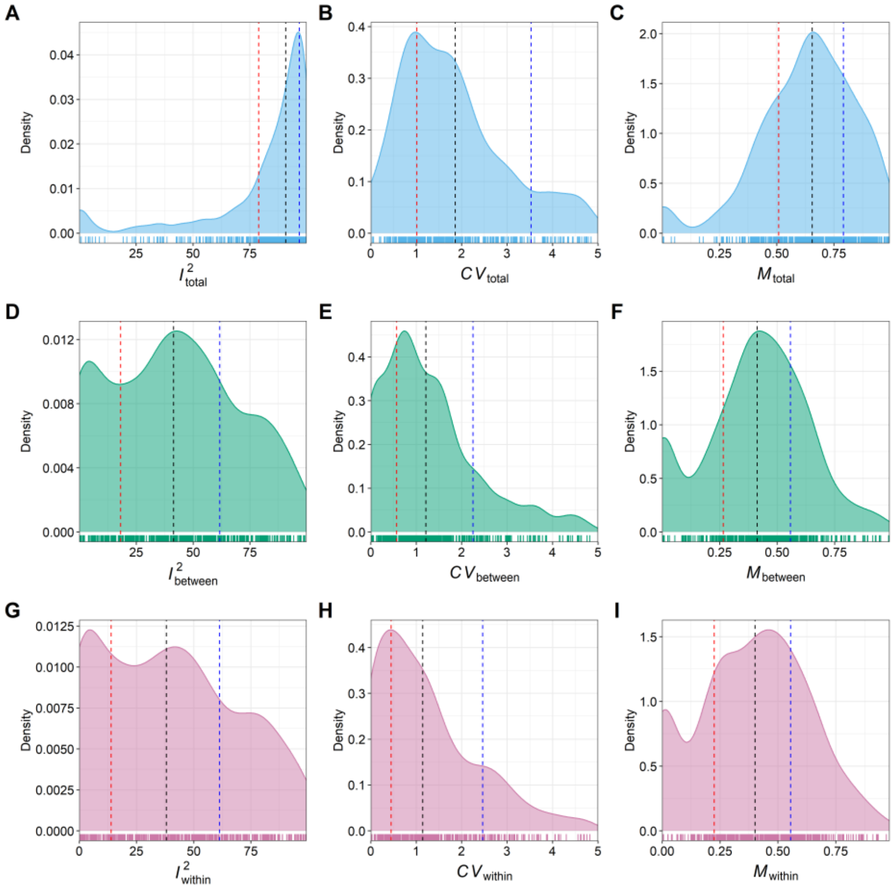


Fig. 3:

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.

When the CV_{total} metric was used to quantify heterogeneity, the calculated 25th, 50th, and 75th percentiles of CV_{total} values were 1.0, 1.8, and 3.5, respectively (Fig. 3). This means that the standard deviation (in this case, heterogeneity) was, on average, 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. In contrast, the distribution of 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).

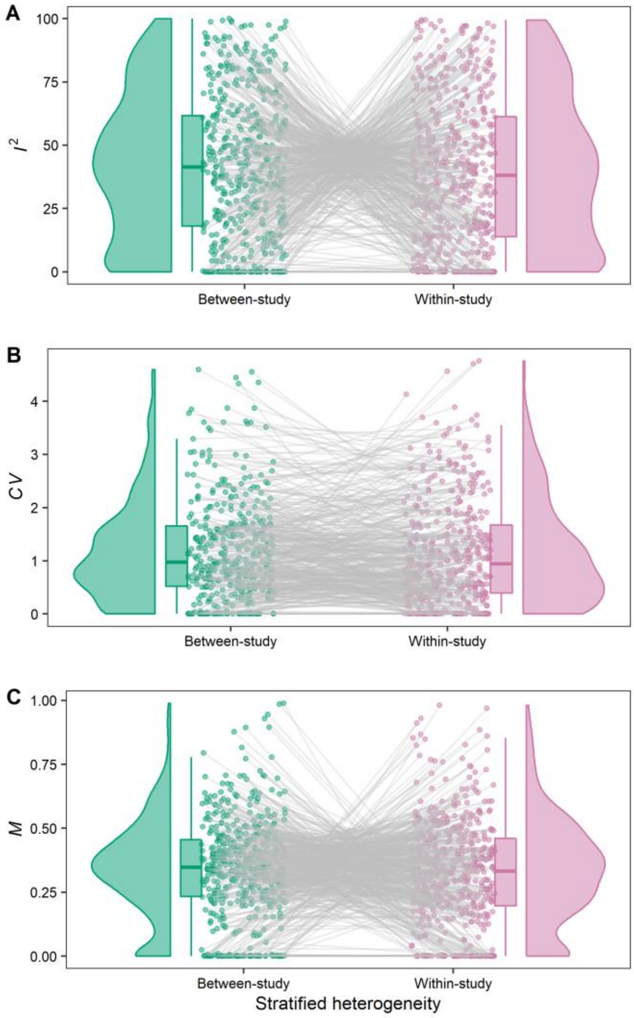


Fig. 4:

Paired comparison of stratified heterogeneity estimates derived 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.

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 [10]), 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. 5). Therefore, cross-meta-analysis (meta-scientific) evidence suggests that the heterogeneity source measure I^2 is not consistent with the 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 **A pluralistic framework**).

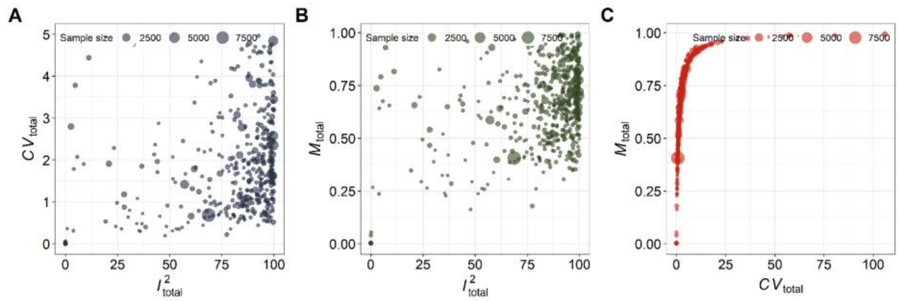


Fig. 5:

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}.

Prediction intervals and predictive distributions: visualising heterogeneity

Prediction intervals (PIs) are underreported but insightful in meta-analytic heterogeneity and generality. Surveys have shown that less than one per cent (1/102) of such studies reported Pis [30]). Pis are derived from the definition of σ_t^2 and provide a range within which a future effect size is predicted to fall with a certain probability [14], often 95% (Fig. 6).

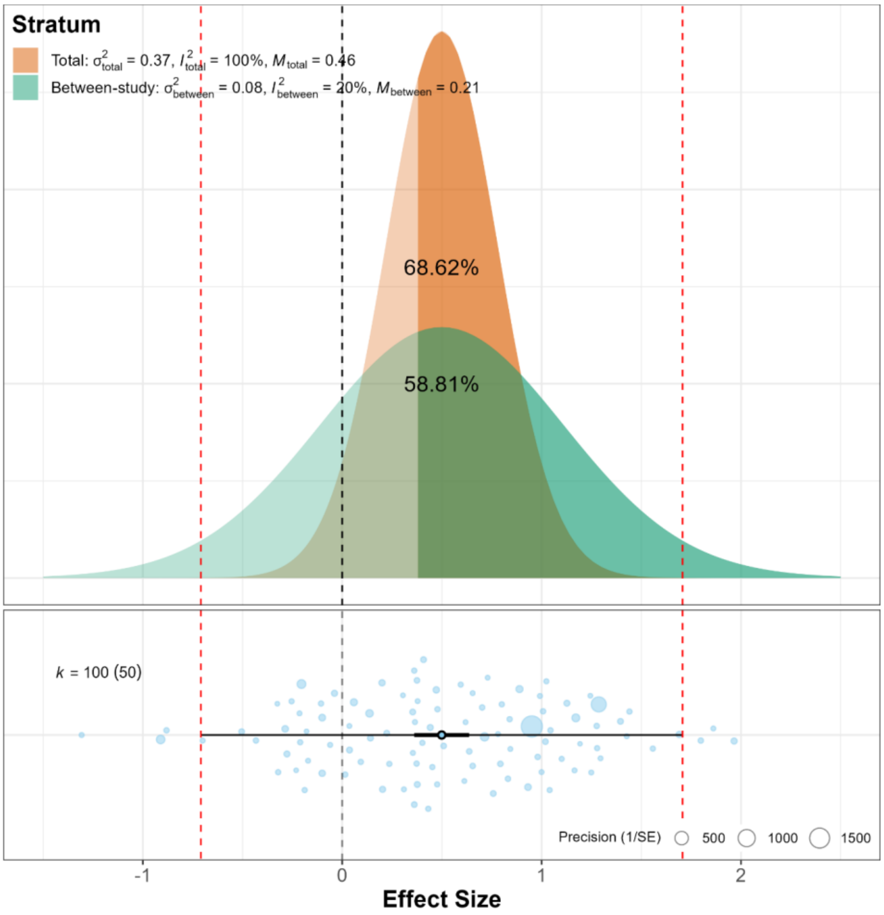


Fig. 6:

Example of how prediction intervals (PIs) combined with ‘prediction distributions’ (PDs) can be used to understand effect size heterogeneity and generality. Effect size data are simulated assuming two effect sizes were collected from a total of $n = 50$ studies, ($k = 100$), with a $\sigma^2_{between} = 0.1$, $\sigma^2_{within} = 0.6$ and an overall meta-analytic mean, u, of 0.5 (https://yefeng0920.github.io/heterogeneity_guide/). Red dashed lines are the upper and lower 95% PI, black dashed line the 'null' effect. The orchard plot [30] displays the overall meta-analytic mean, 95% confidence interval (CI) and 95% PI. The PDs were constructed using *t*-distribution with *k* -1 degrees of freedom, u as location parameter, and total or between-study variance along with sampling variance of around *u* as scale parameter (see Equation 11). The percentage of effect sizes beyond a given threshold (i.e., the lower 95% CI) are provided.

For example, consider a conservation intervention with a mean effect size (SMD) of -0.5 and 95% PI of [-0.2 to -0.8]. This indicates that 95% of future interventions implemented in are predicted to decrease the conservation outcomes of interest by between 0.2 to 0.8 standard deviations. Unlike the point estimate of heterogeneity, such as σ^2_t, PIs offer an interval to inform the extent to which the focal effect can be generalized [31]. Under Equation 1, 95% PIs can be computed by [7]:

$$95\%\text{PI} = \mu \pm t_{0.975}\sqrt{\sigma^2_{between} + \sigma^2_{within} + \text{SE}[\mu]^2}, (11)$$

where $t_{0.975}$ denotes the 97.5*th* percentile of a *t*-distribution (with k−1 degrees of freedom [32], where k is the number of sample size), and SE[μ] denotes the standard error of the mean effect μ.

Despite their usefulness, PIs can create the illusion that all effect sizes within the upper and lower intervals are equally likely (Fig. 6; see also [33]). Therefore, statisticians have emphasised the importance of visualising the probability density to accurately capture the likelihood of each effect size within the intervals [34,35]. By considering the entire distribution of true effects while accounting for statistical noise, the predictive distribution (PD) offers a more holistic measure of heterogeneity and generality. In the Bayesian framework, PDs, known as posterior distributions, are a natural part of the process, but even frequentist approaches can adopt PDs (sometimes referred to as "empirical Bayes") to achieve similar aims. An advantage of the PD is its ability to calculate the probability that a true effect size exceeds a biologically or practically meaningful threshold although determining such a threshold usually requires domain-specific knowledge and expertise. The proportion of true effect sizes above a specific threshold could serve as a measure of evidence strength and generality [16]. Consider a case that 69% of effect sizes representing the efficacy of a conservation intervention are predicted to surpass a threshold value representing a practically significant effect (Fig.6, where we assumed the lower confidence limit representing the threshold). If assuming similar configurations of study contexts in the sampled future cases, we can infer that the intervention will achieve this benefit in 69% of future cases, with strong implications for policymaking.

Modelling additional sources of heterogeneity

In ecological and evolutionary datasets, complexity often arises from the inclusion of diverse species, temporal, and spatial variations [3]. Such complexity 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 [7,9]. 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 [5,36]:

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

where $u_{s[k]}$ 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 [5]:

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

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

We derive the alternative stratified version of measures as follows:

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

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

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

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

Furthermore, the predictive distribution also can be stratified at phylogenetic and non-phylogenetic species-level, which provides a visual means to assess the heterogeneity and generality at these strata.

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. [37]. 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. The data and code for replicating all calculations can be found at https://yefeng0920.github.io/heterogeneity_guide/. Our re-analysis yielded three observations. Firstly, I^2_{total}= 97% exceeded the 75th percentile of the empirically derived heterogeneity distribution (Fig. 7 and Table S1). This suggests a high amount of heterogeneity according to the conventional benchmarks [10]. However, when we employed magnitude metrics to measure heterogeneity, they fell below between 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).

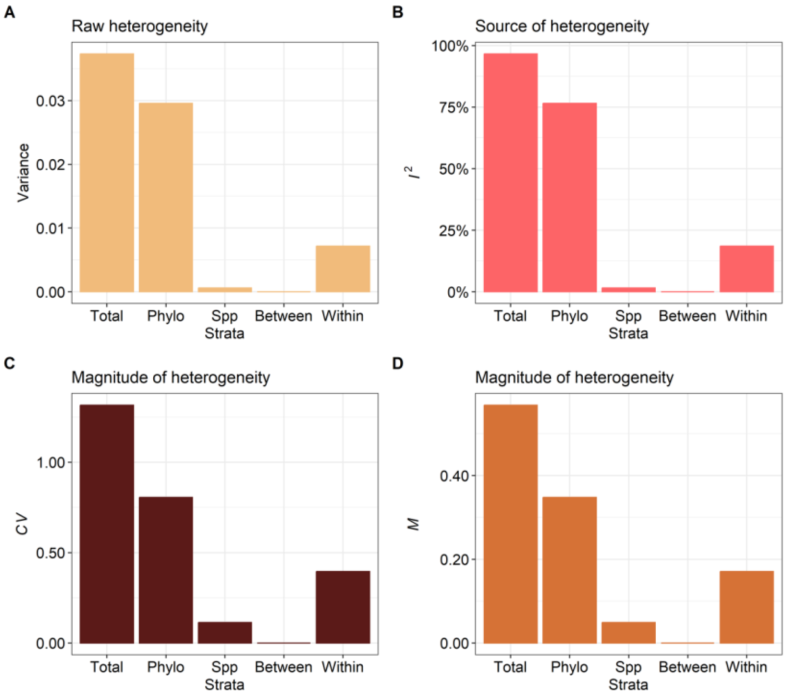


Fig. 7:

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 [10]. 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. 5 and Table S1). However, this amount of heterogeneity was not attributable to the study level but, rather, was mainly explained by the phylogenetic signal ($I^2_{phylogeny}$ = 76%). The stratified version of PD also provided a clearer visual clue that the phylogenetic signal was the primary source of heterogeneity (Fig. 7).

A pluralistic framework

Given that different measures offer distinct insights into heterogeneity and generality (Table 1), we propose adopting a pluralistic framework to comprehensively assess heterogeneity in ecological and evolutionary meta-analyses. Our recommendations are threefold:

(1) Employing multilevel meta-analytic framework: Provided data allow, 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. PI or PD should also be reported to provide a visual identification of the heterogeneity information. These measures provide complementary information, for example, the source, magnitude, and visual clue of heterogeneity (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 **Method**) [39]. 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 crucial 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 (see a case study in **Modelling additional sources of heterogeneity**). However, neither the conventional benchmarks (25, 50, and 75% as small, moderate and high heterogeneity [10]) 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.

We argue that ecologists and evolutionary biologists should treat heterogeneity and the meta-analytic mean effect size with equal importance. We provide a user-friendly tutorial equipped with a set of R functions to streamline the qualification, stratification, and interpretation of heterogeneity https://yefeng0920.github.io/heterogeneity_guide/, empowering ecologists and evolutionary biologists to discern generality.

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.

Commented [SN1]: can we put a disclaimer that the spread could be underestimated - these values could be underestimated if we have publication bias - this is espeically so for CV and M

Should discuss with Shinichi

Commented [YY2R1]: Good point

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 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.	25th, 50th, and 75th percentiles (Fig. 3): 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.
Visual metric	PI & PD	Heterogeneity visual measure. A plausible interval where a new effect size is predicted to fall with a specified level of probability. It can be used to visually diagnose the heterogeneity and generality of the mean effect. For example, a 95% prediction interval (PI) of [-0.2 to -0.8] indicates that 95% range of future effect sizes are expected in studies with similar contexts. The whole predictive distribution (PD) can be used to derive the probability of a newly observed effect being above a biologically meaningful threshold.	Not applicable

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

Methods

Meta-analysis database

The ecological and evolutionary database used in this study were originally compiled by Costello [18], O'Dea [17], and their colleges. They conducted a systematic search 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.

Stratification 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 **Model additional source heterogeneity**). 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 [24,40]. 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_{within}^2, where σ_{within}^2 represents the extent to which within-study effect $\mu_{within[i]}$ deviates from between-study effect $\mu_{between[j]}$[24,40]. 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 [24,40]:

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

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 [40,41].

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

where $\sigma_{total}^2 = \sigma_{between}^2 + \sigma_{within}^2$ is the total variance in effect sizes and $\rho = \sigma_{between}^2/\sigma_{total}^2$ denotes intraclass correlation coefficient. We used the *rma.mv()* function from the *metafor* package [42] 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 ($\sigma_{between}^2$ and σ_{within}^2), 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 ($\sigma_{between}^2$ and σ_{within}^2) by checking their likelihood profiles. The R code for model fitting can be accessed at the https://github.com/Yefeng0920/heterogeneity_ecoevo.

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_{total}^2, an analogue to CV_{total}:

$$CV_{total}^2 = \frac{\sigma_{total}^2}{\mu^2}, (21)$$

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

$$CV_{between}^2 = \frac{\sigma_{between}^2}{\mu^2}, (22)$$

$$CV_{wihtin}^2 = \frac{\sigma_{within}^2}{\mu^2}, (23)$$

Following the same principle, M_{total}^2 can be obtained [11]:

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

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 $\mathrm{logit}(M_{total}^{2}) = \log{(CV_{total}^{2})}$.

Data availability

The data needed to reproduce the analyses and figures are archived GitHub repository https://github.com/Yefeng0920/heterogeneity_ecoevo/tree/main, and will be deposited at Zenodo after acceptance.

Code availability

The scripts needed to reproduce the analyses and figures are archived GitHub repository https://github.com/Yefeng0920/heterogeneity_ecoevo/tree/main, and will be deposited at Zenodo after acceptance.

References

1 Lawton, J. H. Are there general laws in ecology? *Oikos*, 177-192 (1999).

2 Spake, R. *et al.* Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology & Evolution*, 1-11 (2022).

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

4 Martin, P. A. *et al.* Flexible synthesis can deliver more tailored and timely evidence for research and policy. *Proceedings of the National Academy of Sciences* **120**, e2221911120 (2023).

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

6 Noble, D. W. *et al.* Meta-analytic approaches and effect sizes to account for ‘nuisance heterogeneity’in comparative physiology. *Journal of Experimental Biology* **225**, jeb243225 (2022).

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

8 Senior, A. M. *et al.* Heterogeneity in ecological and evolutionary meta ‐ analyses: its magnitude and implications. *Ecology* **97**, 3293-3299 (2016).

9 Nakagawa, S., Yang, Y., Macartney, E. L., Spake, R. & Lagisz, M. Quantitative evidence synthesis: a practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environmental Evidence* **12**, 8, doi:10.1186/s13750-023-00301-6 (2023).

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

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

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

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

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

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

16 Mathur, M. B. & VanderWeele, T. J. New metrics for meta ‐ analyses of heterogeneous effects. *Statistics in Medicine* **38**, 1336-1342 (2019).

17 O'Dea, R. E. *et al.* Preferred reporting items for systematic reviews and meta ‐ analyses in ecology and evolutionary biology: a PRISMA extension. *Biological Reviews* **96**, 1695-1722 (2021).

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

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

20 Yang, Y. *et al.* Robust point and variance estimation for ecological and evolutionary meta-analyses with selective reporting and dependent effect sizes. *EcoEvoRxiv*, doi:https://doi.org/10.32942/X20G6Q (2023).

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

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

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

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

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

26 Nakagawa, S. *et al.* Meta‐analysis of variation: ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution* **6**, 143-152 (2015).

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

28 Yang, Y. *et al.* Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology. *BMC biology* **21**, 1-20 (2023).

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

30 Nakagawa, S. *et al.* The orchard plot: Cultivating a forest plot for use in ecology, evolution, and beyond. *Research Synthesis Methods* **12**, 4-12 (2021).

31 van Aert, R. C., Schmid, C. H., Svensson, D. & Jackson, D. Study specific prediction intervals for random‐effects meta‐analysis: A tutorial: Prediction intervals in meta‐analysis. *Research synthesis methods* **12**, 429-447 (2021).

32 Knapp, G. & Hartung, J. Improved tests for a random effects meta‐regression with a single covariate. *Statistics in medicine* **22**, 2693-2710 (2003).

33 Bishop, J. & Nakagawa, S. Quantifying crop pollinator dependence and its heterogeneity using multi‐level meta‐analysis. *Journal of Applied Ecology* **58**, 1030-1042 (2021).

34 Jackson, C. H. Displaying uncertainty with shading. *The American Statistician* **62**, 340-347 (2008).

35 Barrowman, N. J. & Myers, R. A. Raindrop plots: a new way to display collections of likelihoods and distributions. *The American Statistician* **57**, 268-274 (2003).

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

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

38 Voelkl, B. *et al.* Reproducibility of animal research in light of biological variation. *Nature Reviews Neuroscience*, 1-10 (2020).

39 Raue, A. *et al.* Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923-1929 (2009).

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

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

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

Acknowledgements

YY was funded by the National Natural Science Foundation of China (NO. 32102597). YY, SN, and ML were funded by the Australian Research Council Discovery Grant (DP210100812 & DP230101248). DWAN was supported by an ARC Future Fellowship (FT220100276).

Author contributions

YY: Conceptualization; data curation; formal analysis; investigation; methodology; software; visualization; writing – original draft; writing – review and editing. DWAN: Software; visualization; writing – review and editing. RS: Writing – review and editing. AMS: Writing – review and editing. ML: Visualization; writing – review and editing; funding acquisition; supervision. SN: Conceptualization; investigation; methodology; software; validation; writing – review and editing; funding acquisition; supervision. All authors approved the final manuscript.

Competing interests

All authors declare no competing interests.

Additional information

Supplementary materials will be available at the online version.