A pluralistic framework for measuring, interpreting and decomposing heterogeneity in meta-analysis

Yang, Y. ORCID: https://orcid.org/0000-0002-8610-4016, Noble, D. W. A. ORCID: https://orcid.org/0000-0001-9460-8743, Spake, R. ORCID: https://orcid.org/0000-0003-4671-2225, Senior, A. M. ORCID: https://orcid.org/0000-0001-9805-7280, Lagisz, M. ORCID: https://orcid.org/0000-0002-3993-6127 and Nakagawa, S. ORCID: https://orcid.org/0000-0002-7765-5182 (2025) A pluralistic framework for measuring, interpreting and decomposing heterogeneity in meta-analysis. Methods in Ecology and Evolution, 16 (11). pp. 2710-2725. ISSN 2041-210X doi: 10.1111/2041-210x.70155 Available at https://centaur.reading.ac.uk/124608/

Received: 29 December 2024 | Accepted: 23 June 2025

DOI: 10.1111/2041-210x.70155

Methods in Ecology and Evolution

RESEARCH ARTICLE

A pluralistic framework for measuring, interpreting and decomposing heterogeneity in meta-analysis

Yefeng Yang[1,2] | Daniel W. A. Noble[3] | Rebecca Spake[4] | Alistair M. Senior[5] | Malgorzata Lagisz[1,6] | Shinichi Nakagawa[1,6,7]

[1]Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Australia; [2]College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, Zhejiang, China; [3]Division of Ecology and Evolution, Research School of Biology, The Australian National University, Canberra, Australian Capital Territory, Australia; [4]Ecology and Evolutionary Biology Research Division, School of Biological Sciences, University of Reading, Reading, UK; [5]Charles Perkins Centre, Sydney Precision Data Science Centre, and School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales, Australia; [6]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada and [7]Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, Japan

Correspondence
Yefeng Yang
Email: yefeng.yang1@unsw.edu.au

Funding information
Australian Research Council, Grant/Award Number: DP210100812 and DP230101248; Canada Excellence Research Chairs, Government of Canada, Grant/Award Number: CERC-2022-00074; ARC Future Fellowship, Grant/Award Number: FT220100276 and FT230100240

Handling Editor: Aaron Ellison

Abstract

1. Measuring heterogeneity, or inconsistency, among effect sizes is a crucial step for interpreting meta-analytic evidence across diverse taxonomic groups and spatiotemporal contexts. However, ecologists and evolutionary biologists often interpret overall mean effects (mean population effects) as consistent across contexts, either explicitly or implicitly, without properly quantifying and interpreting heterogeneity.
2. Here, we present a pluralistic approach that aims to quantify heterogeneity by introducing complementary metrics, each of which decomposes heterogeneity into within-study, between-study and between-species (species and phylogenetic) variances. These metrics include the traditional I^2 (variance-standardized metric), the newly derived coefficient of variation for heterogeneity (*CVH* family; mean-standardized metric), the second-order coefficient of variation (*M* family; variance–mean-standardized metric) and their stratified variants.
3. To demonstrate the benefits of the combined use of these measures, we synthesize heterogeneity estimates from 512 ecological and evolutionary meta-analyses. We show that total heterogeneity (variance of true effects) is, on average, 10 times larger than statistical noise (sampling error 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 $CVH = 1.8$ and $M = 0.6$), indicating substantial variation among studies within a meta-analysis. Moreover, different effect size types yield different values of heterogeneity metrics because they are inherently influenced by statistical properties of their effect size estimators. As such, comparisons of heterogeneity across effect size types should be made with caution, albeit the proposed heterogeneity metrics are unit-free.

4. Our large-scale synthesis also provides new benchmarks for the interpretation of heterogeneity and recommendations on how to quantify and report heterogeneity. New extensions for stratifying heterogeneity metrics will clarify our understanding of the generalisability, and at what level of meta-analytic effects in ecology and evolution.

KEYWORDS
context dependence, effect size, heterogeneity, linear models, meta-analysis, mixed effects model

1 | INTRODUCTION

Meta-analytic modelling is widely used to test ecological and evolutionary hypotheses, which can be important in informing conservation and environmental policy (Gurevitch et al., 2018). Three critical steps are necessary. First, an estimate of an overall mean effect characterizes the magnitude of a focal effect of interest (Nakagawa & Santos, 2012; Yang, Lagisz, et al., 2024). Second, a measure that quantifies the inconsistency among study findings, the 'heterogeneity' among true effect sizes, is estimated to contextualize study findings. Finally, effect modifiers or moderator variables that are hypothesized to explain variation in effect sizes—and how much of it is identified (context-specific effects; Nakagawa & Santos, 2012). Crucially, heterogeneity indicates 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. Without quantifying heterogeneity, it is not possible to properly interpret both the overall trends and context-specific effects (Senior et al., 2016; Spake et al., 2022).

While meta-analyses of a collection of studies using similar protocols for single species allow for clearer interpretations, the interpretation of average population effects across diverse taxonomic groups and spatiotemporal contexts can be difficult. 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 across the contexts when the meta-analytic evidence base accounts for informative effect modifiers, leading to a low amount of 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 in practice. Indeed, surveys have revealed that heterogeneity statistics are not routinely reported (Nakagawa et al., 2023; Senior et al., 2016; Yang et al., 2022).

Currently, measuring and interpreting meta-analytic heterogeneity is challenging for two major reasons. First, no single heterogeneity metric provides a holistic interpretation of inconsistency among study findings (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 error 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 Figure 1; Borenstein et al., 2017; IntHout et al., 2016; Rücker et al., 2008). Second, meta-analytic practice typically focuses on estimating total heterogeneity only (Nakagawa & Santos, 2012), despite the hierarchical nature of real biological data structures (Nakagawa et al., 2023; Noble et al., 2022). 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 that drive context dependence (Nakagawa & Santos, 2012). For example, in a multi-taxon meta-analysis, if stratification of studies by species yields low heterogeneity at the taxon level, the focal effect can still be generalizable across taxon (Figure 2). This is so, even if the total heterogeneity remains high (Senior et al., 2016).

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 applies 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. Costello & Fox, 2022; O'Dea et al., 2021). We also present meta-scientific evidence on (in)congruence between different heterogeneity metrics and outline approaches for developing useful extensions of heterogeneity quantification for phylogenetic multilevel meta-analyses. The replication materials for this study are available on the GitHub repository (https://github.com/Yefeng0920/heterogeneity_benchmark) and Zenodo (Yang, 2025). 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

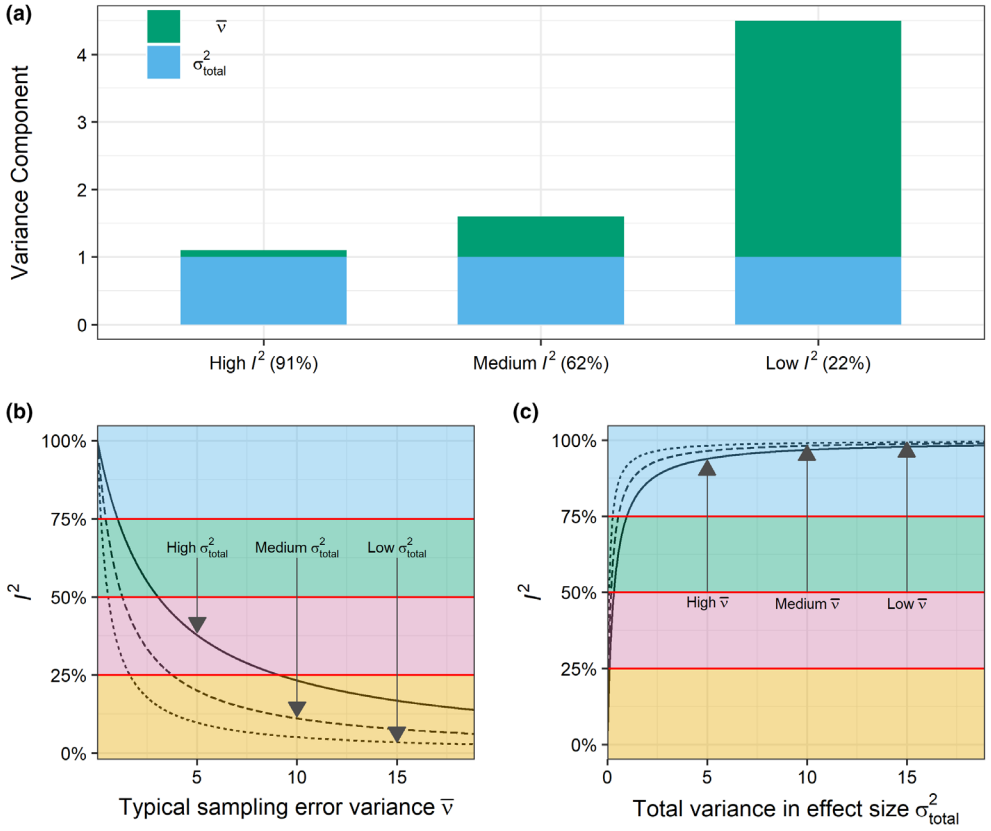

FIGURE 1 The interpretation of total I^2 can be ambiguous and can lead to incorrect conclusions about the magnitude of heterogeneity. (a) The value of the total I^2 is dependent on sampling error variances. (b) A large estimated total I^2 value could be due to small 'typical' sampling error variances $\bar{v}$ (Equation 3). (c) In contrast, 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. 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.

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.

2 | METHODS

2.1 | Database

The ecological and evolutionary databases used in this study were originally compiled by Costello and Fox (2022) and O'Dea et al. (2021). For more information on data collection, see the relevant data sources (Costello & Fox, 2022; O'Dea et al., 2021). After de-duplicating, our database included 522 meta-analytic datasets (Yang, 2025). We dropped meta-analysis datasets that could not achieve convergence when fitted to the multilevel model. Table S1 reports a descriptive summary of these datasets that were excluded due to model convergence issues. Convergence could not be reached for nine meta-analytic 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. Each meta-analysis dataset included, on average, 240 effect size estimates (first quartile = 30, median = 68, third quartile = 201) from 40 studies (first quartile = 12, median = 24, third quartile = 49).

2.2 | Stratifying heterogeneity using a 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 (i.e. between- and within-study levels; Equation 1). Ecological and evolutionary meta-analyses typically report around six effect size estimates per study (median). However, traditional random-effects meta-analytic approaches do not account for heterogeneity driven by such data stratification (Nakagawa et al., 2023; Noble et al., 2022; Yang et al., 2022), and multilevel meta-analysis is required to model heterogeneity at different strata or multilevel in a meta-analysis (see Appendix S1 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

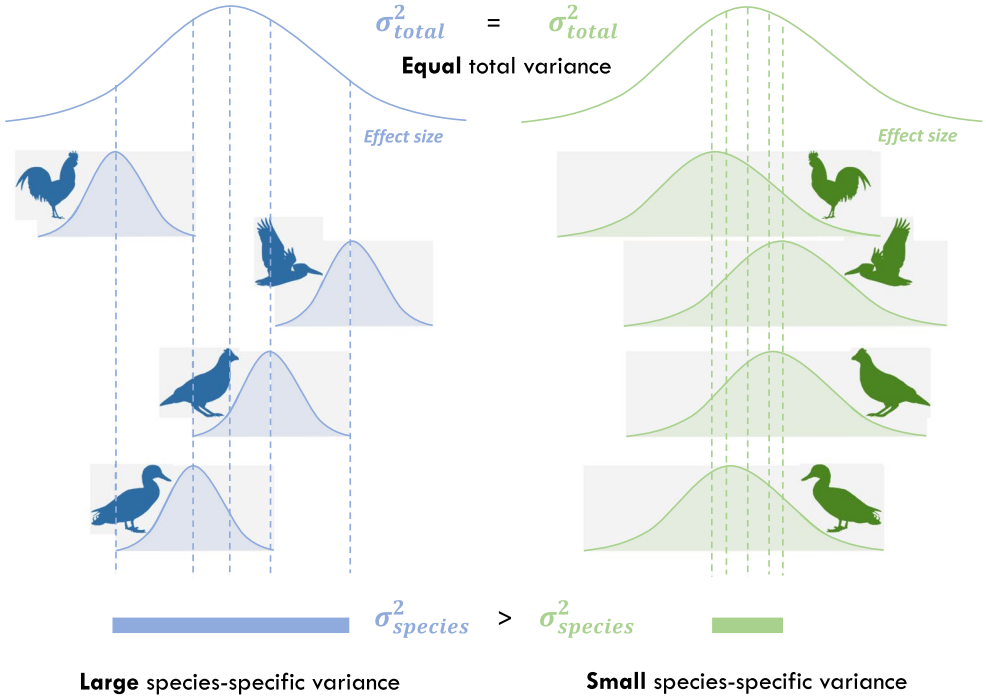

FIGURE 2 A cross-taxa meta-analysis with a high total variance can have a small amount of species-level heterogeneity. It is still possible that the focal effect will be generalizable at the species level. The circles represent the replicated species-specific effects. The red dashed lines denote the meta-analytic mean effects. See a real example in Extended strategies: Non-phylogenetic and phylogenetic species-level heterogeneity and generality.

meta-analytic overall mean effect size μ (overall mean of an outcome of interest), random effects at two strata (i.e. between- and within-study levels) and sampling error effect:

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

The typical assumptions for Equation (1) are as follows: (i) between-study-level random effect $u_{between[j]}$ 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 $e_{[i]} \sim \mathcal{N}(0, v_{[i]})$. 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). Note that in the context of the traditional random-effects model, the between-study variance, often termed τ^2, is treated as the σ^2_{total}. In contrast, a multilevel model (essentially a random-effects model with multiple random effects) treats between-study variance as one of the components of the σ^2_{total}. Therefore, $\tau^2 = \sigma^2_{between} = \sigma^2_{total}$, when $\sigma^2_{within} = 0$.

Statistical analyses were carried out using R 4.0.3 computing platform (R Core Team, 2020). We used the *rma.mv()* function from the *metafor* package (v4.7.53; Viechtbauer, 2010) to fit all 512 meta-analysis datasets to the multilevel meta-analytic model (Equation 1). We employed restricted maximum likelihood REML (embedded in *metafor* package) 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. 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 on the website (see Supporting Information; https://yefeng0920.github.io/heterogeneity_guide/). In the following sections, we will elaborate on how to use Equation (1) to stratify heterogeneity information for different metrics.

2.3 | Complementary measures of heterogeneity

2.3.1 | Unstandardized 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 (hereafter referred to as 'raw heterogeneity'). The square roots of σ^2_{total}, $\sigma^2_{between}$ and σ^2_{within} represent the standard deviation of the true effect size and can also be used as a direct measure of absolute heterogeneity. In Equation (1), the variance of the observed effects ($Var[ES_{[i]}]$) is the

sum of the sampling error variance and the true effect variance (σ^2_{total}). In meta-analyses with infinite sample sizes, $\mathrm{Var}\left[\mathrm{ES}_{[i]}\right]$ is larger than σ^2_{total}. Importantly, Equation (1) provides a general way to partition σ^2_{total} into 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 (Figure 2). For example, low variation among species implies effects are similar, on average, across species. Nonetheless, relying solely on absolute variance does not provide practical intuition regarding the magnitude of heterogeneity. For example, in a meta-analysis with $\sigma^2_{total}=1$, it is unclear whether this amount of variance is large and meaningful because absolute variance is not unitless and comparable across effect size statistics. Importantly, interpreting σ^2_{total} in context is crucial because its magnitude depends on the research field, study designs and measurement scales. A proper contextual interpretation requires a thorough understanding of the topic, including typical effect size ranges, study characteristics and sources of variability. However, if contextual interpretation is unclear or difficult due to limited subject knowledge, researchers can resort to empirical benchmarks, such as median or quartile σ^2_{total} values from similar meta-analyses (see Section 3.2). These benchmarks provide a reference point, helping to assess whether observed heterogeneity is typical, moderate or extreme relative to comparable syntheses. While empirical benchmarks can be a practical guide, they should complement, not replace, efforts to understand heterogeneity in the specific context of the research question.

2.3.2 | Variance-standardized heterogeneity metrics

The heterogeneity index, I^2 has emerged as the most popular heterogeneity 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 sampling error variance (Higgins & Thompson, 2002). The total I^2 (denoted as I^2_{total}) can be computed by dividing the variance in the true effects (σ^2_{total}) by the variance in the observed effects ($\mathrm{Var}\left[\mathrm{ES}_{[i]}\right]$). Therefore, I^2_{total} is given by

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

where $\bar{v}$ represents the 'typical' sampling error variance, representing the average level of sampling error variance. $\bar{v}$ can be computed using different estimators (Cheung, 2014; Takkouche et al., 1999), with the common one being (Higgins & Thompson, 2002):

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

where k denotes the number of observations (in this case, effect size estimates). Within the multilevel modelling framework, the total I^2 can be stratified, for example, by estimating I^2 at between-study ($I^2_{between}$) and within-study (I^2_{within}) levels (Cheung, 2014; Nakagawa & Santos, 2012):

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

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

However, as mentioned earlier, large I^2 values do not necessarily imply a practically relevant amount of heterogeneity (see Figure 1; also see a case study in 'Extended strategies: Non-phylogenetic and phylogenetic species-level heterogeneity and generality'). Stratified I^2 metrics range from 0 to 1 (or can be rescaled to a percentage ranging from 0 to 100 percent), 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 0.9 means within-study variation accounts for 90% of I^2_{total}, 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 $\log\left(\frac{\sigma^2}{\bar{v}}\right)=\mathrm{logit}(I^2)$), which represents heterogeneity as a proportion of the sampling error variance.

2.3.3 | Mean-standardized heterogeneity metrics

Evolutionary biologists and behavioural ecologists are familiar with variance-scaled metrics such as heritability (h^2) and repeatability (R), which are statistically comparable to the variance-scaled heterogeneity index, I^2. Less commonly used but equally relevant are 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 et al., 2011). Here, we introduce a mean-scaled heterogeneity metric, $CVH2_{total}$ ('H' and '2' denoting 'heterogeneity' and 'squared version', respectively) that can be used in meta-analysis, which standardizes heterogeneity by comparing the variance of true effects (σ^2_{total}) to the square of the overall mean effect size (μ^2) (Takkouche et al., 1999):

$$CVH2_{total}=\frac{\sigma^2_{total}}{\mu^2}. \quad (6)$$

$CVH2_{total}$ can be easily interpreted as it expresses heterogeneity as a proportion of the overall mean effect size, or as a percentage when multiplied by 100. A value of $CVH2_{total}=1$ indicates that the heterogeneity (variance among true effects) equals the overall mean effect size. Assuming a normal distribution this means ~16%

TABLE 1 Rule-of-thumb and empirically derived benchmarks for the interpretation of heterogeneity based on I^2, *CVH2* and *M2*.

	Tentative interpretation benchmarks			
	Rule-of-thumb		Empirically-derived[a]	
Metric	Category	Range	Percentile	Range
I^2	Very small	0 to 0.25	0th to 25th	0 to 0.79
	Small	0.25 to 0.50	25th to 50th	0.79 to 0.91
	Moderate	0.50 to 0.75	50th to 75th	0.91 to 0.97
	Large	0.75 to 1	75th to 100th	0.97 to 1
CVH2	Very small	0 to 0.04	0th to 25th	0 to 1.03
	Small	0.04 to 0.19	25th to 50th	1.03 to 3.45
	Moderate	0.19 to 0.56	50th to 75th	3.45 to 12.43
	Large	0.56 to ∞	75th to 100th	12.43 to ∞
M2	Very small	0 to 0.04	0th to 25th	0 to 0.51
	Small	0.04 to 0.16	25th to 50th	0.51 to 0.78
	Moderate	0.16 to 0.36	50th to 75th	0.78 to 0.93
	Large	0.36 to 1	75th to 100th	0.93 to 1

Note: Table S3 provides empirically derived benchmarks for the full set of standardized heterogeneity metrics. The rule-of-thumb was retrieved from the literature, with slight modifications (Higgins et al., 2003; Kvålseth, 2017). Empirically derived interpretation benchmarks are proposed based on the empirical distribution of different heterogeneity measures. Table 2 provides the empirically derived benchmarks corresponding to the commonly used effect size measures (e.g. Cohen's *d*). Given the differences between different effect size measures, we recommend using effect size type-specific benchmarks (but see the limitations of using empirically derived benchmarks in Section 3.2). Definitions of heterogeneity measures can be found in both the main text and the Appendix S1. For simplicity, the subscript for each heterogeneity measure was removed in Table 1. The precise percentile range in which the heterogeneity estimates for a particular meta-analysis fall can be obtained via the R helper function *het_interpret()*.

[a]The distributions and percentiles could be underestimated if publication bias existed. While the existing technique allows for publication bias to be taken into account to obtain bias-corrected estimates of the population mean effect (Yang, Lagisz, et al., 2024), there is not yet any method to obtain bias-corrected estimates of heterogeneity.

of effects would have opposite sign to the overall mean effect (Figure S1). To assist with interpretation, we provide rule-of-thumb and empirically derived benchmarks to classify heterogeneity as 'very small', 'small', 'medium' or 'large' (Tables 1 and 2). In addition, we provide an R helper function (*het_interpret()*) that can help determine the percentile range in which the heterogeneity estimates for a particular meta-analysis fall, based on the heterogeneity distribution of the published meta-analyses.

For a more precise breakdown of heterogeneity, we propose two variants of $CVH2_{total}$ under the multilevel model framework (Equation 1). We express the between-study, $CVH2_{between}$, and within-study, $CVH2_{within}$, versions of $CVH2_{total}$ as follows:

$$CVH2_{between} = \frac{\sigma^2_{between}}{\mu^2}, \quad (7)$$

$$CVH2_{within} = \frac{\sigma^2_{within}}{\mu^2}. \quad (8)$$

These variants quantify between- and within-study heterogeneity relative to the effect being measured. Additionally, we provide mean-standardized metrics based on standard deviation (e.g. σ_{within}) rather than variances (e.g. σ^2_{within}), $CVH1_{total}$, $CVH1_{between}$ and $CVH1_{within}$ (see Appendix S1). To estimate $CVH2_{total}$ and its two variates, we suggest using the maximum likelihood estimates for $\sigma^2_{between}$, σ^2_{within} and μ derived from Equation (1), and substitute them into Equations (6–8). For simplicity, throughout the paper, we use population parameters (e.g. $\sigma^2_{between}$, σ^2_{within} and μ) and their estimators (e.g. $\overline{\sigma}^2_{between}$, $\overline{\sigma}^2_{within}$ and $\overline{\mu}$) interchangeably. Notably, these mean-scaled variance metrics have the limitation of becoming arbitrarily large as the magnitude of overall mean effect μ approaches zero (Kvålseth, 2017; Lobry et al., 2023).

2.3.4 | Variance–mean-standardized heterogeneity metrics

To remedy the limitations of I^2_{total} and $CVH2_{total}$ as illustrated above, we introduce a more robust heterogeneity measure, $M2_{total}$, which combines the strengths of mean-scaled and variance-scaled metrics (Cairns & Prendergast, 2022; Kvålseth, 2017):

$$M2_{total} = \frac{\sigma^2_{total}}{\sigma^2_{total} + \mu^2}. \quad (9)$$

TABLE 2 Summary of heterogeneity measures and their stratified counterparts. SMD denotes standardized mean difference.

Types	Metrics	Interpretation and examples	Empirically derived benchmark[a]
Test statistic	Q	Null-hypothesis test. Statistical test of heterogeneity in effect sizes	Not applicable
Unstandardisation	σ^2 family	Absolute magnitude measure of heterogeneity. Variance (square of standard deviation) of the meta-analytic overall mean effect (σ^2_{total}) and its stratification in between- and within-study contexts ($\sigma^2_{between}$ and σ^2_{within}).	25th, 50th and 75th percentiles (Figure S4): 0.54, 1.25 and 3.03 for SMD; 0.11, 0.27 and 0.57 for lnRR; 0.06, 0.12 and 0.25 for *Zr*; 1.04, 1.20 and 2.51 for the 2-by-2 table; 0.01, 0.04 and 0.27 for uncommon measures. The percentiles of typical sampling variance $\bar{v}$ are reported at Figure S5.
Variance-standardization	I^2 family	Heterogeneity source measure. Proportion of variance not due to sampling error variance. It measures the source of heterogeneity. For example, $I^2_{total} = 95\,\%$ denotes that 95% of variation is the result of heterogeneity (i.e. differences in contexts). $I^2_{between} = 0.8$ and $I^2_{within} = 0.15$ indicates 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 (Figure 3): 0.78, 0.89 and 0.96 for SMD; 0.88, 0.95 and 0.99 for lnRR; 0.73, 0.87 and 0.95 for *Zr*; 0.71, 0.73 and 0.89 for the 2-by-2 table; 0.74, 0.91 and 0.98 for uncommon measures.
Mean-standardization	*CVH* family	Heterogeneity magnitude measure, including *CVH*1 and *CVH*2. Variance is expressed as the proportion of the mean effect. It is the measure of the magnitude of heterogeneity in the context of the mean effect. For example, $CVH2_{total} = 1.5$, $CVH2_{between} = 0.8$ and $CVH2_{within} = 0.5$ denotes that total, between- and within-study variance are 150%, 80% and 50% of the mean effect.	25th, 50th and 75th percentiles for *CVH*2 (and *CVH*1): 1.1 (1.05), 3.94 (1.98) and 15.4 (3.93) for SMD; 1.36 (1.16), 3.76 (1.94) and 12.1 (3.48) for lnRR; 0.67 (0.82), 2.77 (1.66) and 8.54 (2.92) for *Zr*; 1.57 (1.21), 4.96 (2.19) and 7.04 (2.65) for the 2-by-2 table; 0.47 (0.69), 1.22 (1.11) and 1.7 (1.3) for uncommon measures.
Variance–mean-standardization	*M* family	Heterogeneity magnitude measure, including *M*1 and *M*2. Variance is expressed as the proportion of the mean effect and a transformation of *CVH* family designed with better properties. It is the measure of the magnitude of heterogeneity in the context of the mean effect.	25th, 50th and 75th percentiles *M*2 (and *M*1): 0.52 (0.51), 0.8 (0.66) and 0.94 (0.8) for SMD; 0.58 (0.54), 0.79 (0.66) and 0.78 for lnRR; 0.4 (0.45), 0.73 (0.62) and 0.9 (0.75) for *Zr*; 0.57 (0.54), 0.82 (0.68) and 0.88 (0.73) for the 2-by-2 table; 0.32 (0.41), 0.55 (0.52) and 0.62 (0.56) for uncommon measures.

Note: 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 and log risk ratio. Uncommon measures represent less frequently used effect size measures, such as raw mean difference and regression coefficients. For simplicity, the subscript for each heterogeneity measure was removed in Table 2.

[a]The distributions and percentiles could be underestimated if publication bias existed. While the existing technique allows for publication bias to be taken into account to obtain bias-corrected estimates of the population mean effect (Yang, Lagisz, et al., 2024), there is not yet any method to obtain bias-corrected estimates of heterogeneity.

We also propose stratified versions of $M2_{total}$ for between-study ($M2_{between}$) and within-study ($M2_{within}$) heterogeneity, allowing for a more precise quantification of heterogeneity at specific strata:

$$M2_{between} = \frac{\sigma^2_{between}}{\sigma^2_{total} + \mu^2}, \quad (10)$$

$$M2_{within} = \frac{\sigma^2_{within}}{\sigma^2_{total} + \mu^2}. \quad (11)$$

Similar to $CVH2_{total}$, $M2_{total}$ and its stratified variants provide a standardized measure of heterogeneity relative to the overall mean effect size. Importantly, $M2_{total}$ offers the advantage of being bounded between 0 and 1, making interpretation more intuitive and simpler. For example, $\sigma_{total} = 0$ leads to $M2_{total} = 0$, 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'). Conversely, a value near 1 suggests that heterogeneity is maximized relative to the overall mean effect size. Additionally, $M2_{total}$ can be transformed into a coefficient of variation by applying the logit transformation: $\text{logit}(M2_{total}) = 2\log(CVH2)$. Unlike $CVH2_{total}$, $M2_{total}$ and its stratified variants avoid the problem of over-inflation when the magnitude of overall mean effect μ approaches zero, making it a more robust and reliable measure of heterogeneity.

In the Appendix S1, we describe additional metrics, $M1_{total}$, $M1_{between}$ and $M1_{within}$, where the squared terms in the numerator and denominator are replaced by their square roots. In the

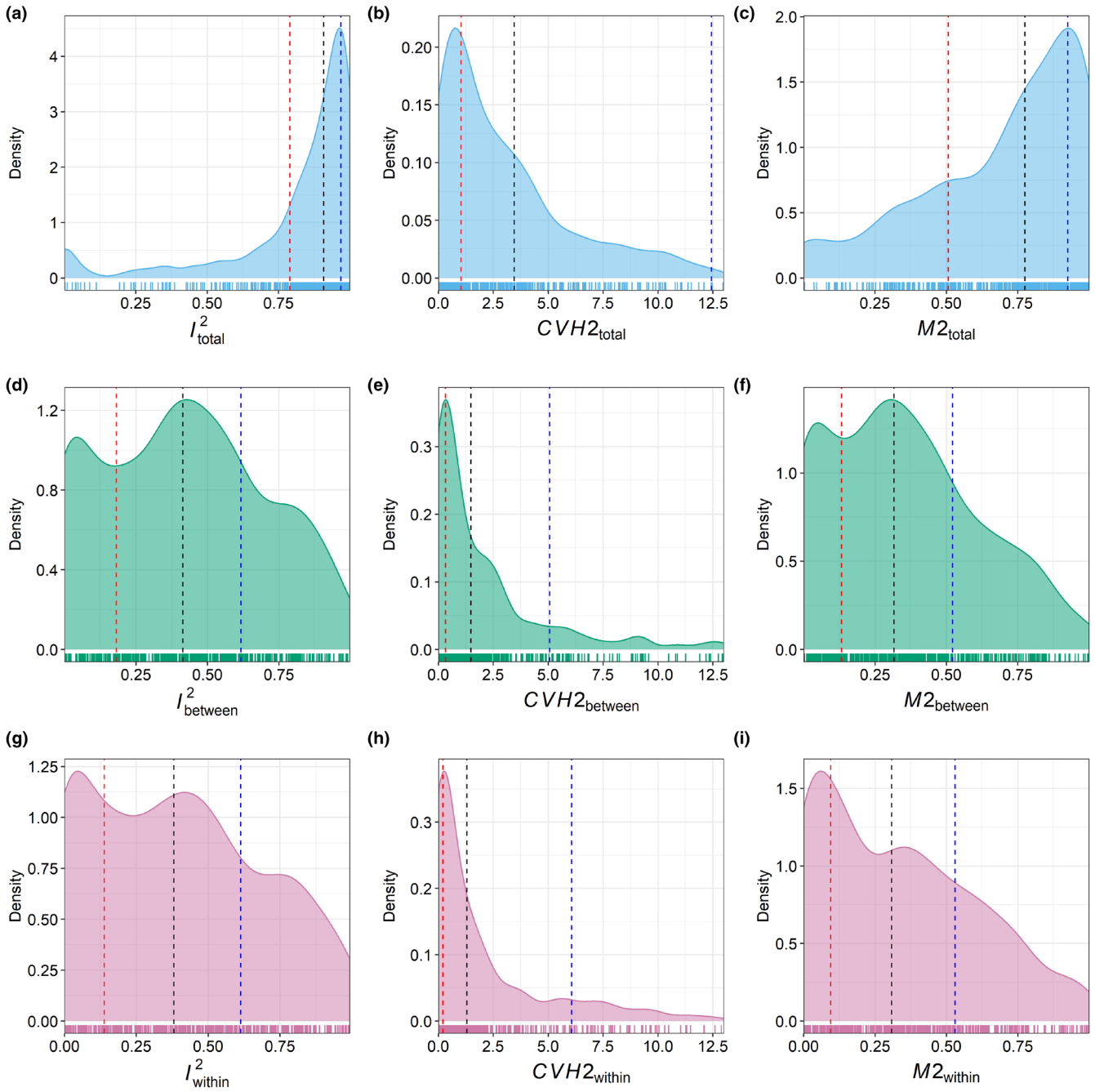

FIGURE 3 The distribution of heterogeneity estimates derived from 512 meta-analyses was systematically assessed using multiple measures and stratified across different strata. Total heterogeneity measures (a–c): I^2_{total}, $CVH2_{total}$ and $M2_{total}$. Between-study heterogeneity measures (d, e): $I^2_{between}$, $CVH2_{between}$ and $M2_{between}$. Within-study heterogeneity measures (g–i): I^2_{within}, $CVH2_{within}$ and $M2_{within}$. Three dashed lines correspond to the 25th, 50th and 75th percentiles, respectively. In panels (b, e and h), the *CVH2* was truncated at five for figure clarity, as very large *CVH2* values can be challenging to interpret when the meta-analytic mean effect is small. For example, the maximum *CVH2* observed in the 512 meta-analyses was 106, which was inflated by a small meta-analytic mean effect of 0.03. For unstandardized (raw) heterogeneity and typical sampling error variance, please refer to Figures S4 and S5. The density of heterogeneity distribution was based on Gaussian kernel density estimation. The degree of smoothing (bandwidth) was determined using a rule-of-thumb method (Heidenreich et al., 2013), which uses 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power. Given that density estimates are sensitive to bandwidth selection (Pick et al., 2023), we also provided the histograms corresponding to panels (a–i) (Figure S2).

statistical literature (Kvålseth, 2017), $M1_{total}$ and $M2_{total}$ are known as second-order coefficients of variation, derived from the ratio of second-order moments. Statisticians interpret these measures in terms of Euclidean distances (a measure of deviation) between true effect sizes and the overall mean effect relative to the distance between true effect sizes and the origin (see geometric

formulation in Kvålseth, 2017). For example, a value of $M2_{total} = 0.5$ means that, in an *n*-dimensional space, the distance (deviation) between a collection of effect sizes and the overall mean effect is 50% of the distance (deviation) to the origin. Although this distance-based interpretation feels unfamiliar, it is worth noting that standard deviation and variance—the most commonly used measures of dispersion—are also based on distances. Standard deviation represents the 'standard' or 'typical' distance (deviation) of a value from the mean value, while variance measures the average of the squared distances from the mean. To further aid in the interpretation of these variance–mean-standardized metrics, we provide both rule-of-thumb and empirically derived benchmarks to categorize heterogeneity as 'small', 'medium' or 'large' (Tables 1 and 2, and R help function *het_interpret()*).

3 | RESULTS AND DISCUSSION

3.1 | Empirical patterns of heterogeneity and implications for the generalizability of the meta-analytic effects

3.1.1 | Source of heterogeneity

To examine the magnitude and sources of heterogeneity across the 512 ecological and evolutionary meta-analyses, we first used the variance-standardized metric I^2, which quantifies the proportion of total observed variance attributable to variation in true effects (as opposed to sampling error). Across the full dataset, which includes meta-analyses using different effect size metrics, the 25th, 50th and 75th percentiles of total heterogeneity (I^2_{total}) were 0.79, 0.91 and 0.97 of I^2_{total}, respectively (Figure 3; Table 1). Importantly, however, these summary values should not be interpreted as universal benchmarks that apply across effect size metrics. The magnitude of I^2_{total}, the raw heterogeneity measure (variance of true effects; σ^2) and the average sampling error variance ($\bar{v}$) are inherently influenced by the scale and statistical properties of the effect size metric used. Indeed, differences emerged when stratifying by effect size type (Table 2): 0.78 (25th), 0.89 (50th) and 0.96 (75th) for standardized mean difference (SMD), 0.88, 0.95 and 0.99 for log response ratio (lnRR), and 0.73, 0.87 and 0.95 for Fisher's z-transformed correlation coefficient (Zr). These differences stem from variation in the magnitude of σ^2 an $\bar{v}$ across effect size types.

The observed distribution of I^2_{total} contrast with the conventional thresholds for interpreting I^2, which typically categorize heterogeneity as small, moderate or high at 0.25, 0.50 and 0.75 of I^2_{total} (Higgins et al., 2003), respectively. Thus, on average (50th percentile), 91% of the variance in effect sizes can be attributed to the 'true' biological or methodological differences in research contexts, and may therefore be explainable using appropriate predictor variables (i.e., moderators). It also indicates that the variance in true effect sizes is 10 times larger than the typical sampling error variance ($\frac{\sigma^2}{\bar{v}} = \frac{I^2}{(1-I^2)} = 10$; see Figures S4 and S5 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 (Figure 3). There was no consistent trend suggesting neither type of stratified heterogeneity consistently outweighed the other across the 512 meta-analyses (Figure 3). Intriguingly, 47% (242 out of 512) of the meta-analyses exhibited smaller between-study level heterogeneity than within-study level heterogeneity ($I^2_{between} < I^2_{within}$; Figure 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.

We note that the above results were drawn from fitting a generic model to meta-analytic datasets without contextualizing any specific ecological and evolutionary topics. Therefore, the above conclusion about heterogeneity accounting for 91% of total variance does not necessarily imply that a given meta-analysis included in our dataset exhibits a high level of heterogeneity and thus a low level of generalizability, although on average this is the case. In contrast, the degree of heterogeneity and generalizability of a specific meta-analysis is linked to the characteristics (e.g. taxonomic coverage, outcomes, study design) of primary studies included in the meta-analysis. Ecologists and evolutionary biologists are encouraged to identify sources of heterogeneity specific to their meta-analyses, testing relevant hypotheses and drawing conclusions about the generalizability of a given effect of interest. For example, telomere length measurements are affected by the laboratory assay (Monaghan et al., 2018; Salmón & Burraco, 2022), with the in-gel hybridization-based TRF method yielding different measurements compared to Southern blot-based TRF and qPCR methods (Chik et al., 2022; Remot et al., 2022). Meta-analysing results of primary studies using different laboratory assays would naturally lead to a high amount of heterogeneity, resulting in low generalizability across studies. However, if the laboratory assay could account for heterogeneity driven by the method of choice (Remot et al., 2022), generalizability could then be concluded as high when conditioned on the method used.

3.1.2 | Magnitude of heterogeneity

When the mean-standardized metric $CVH2_{total}$ was used to quantify the magnitude of heterogeneity, the calculated 25th, 50th and 75th percentiles of $CVH2_{total}$ values were 1.0, 1.8 and 3.5, respectively (Figure 3). Therefore, the variance (raw heterogeneity) was, on average (50th percentile), nearly twice that of the square of the overall mean effect. The distributions of both $CVH2_{total}$ and its stratified versions, $CVH2_{between}$, and $CVH2_{within}$, displayed a right-skewed pattern with a single-mode (Figure 3). In contrast, the distribution of the mean–variance-standardized metric $M2_{total}$ exhibited a more symmetrical pattern, with the 25th, 50th and 75th percentiles of $M2_{total}$ values being 0.5, 0.6 and 0.8, respectively (Figure 3), albeit with a minor peak around zero.

Notably, stratification analysis revealed that $MH2_{between}$ and $MH2_{within}$ had patterns similar to those observed for

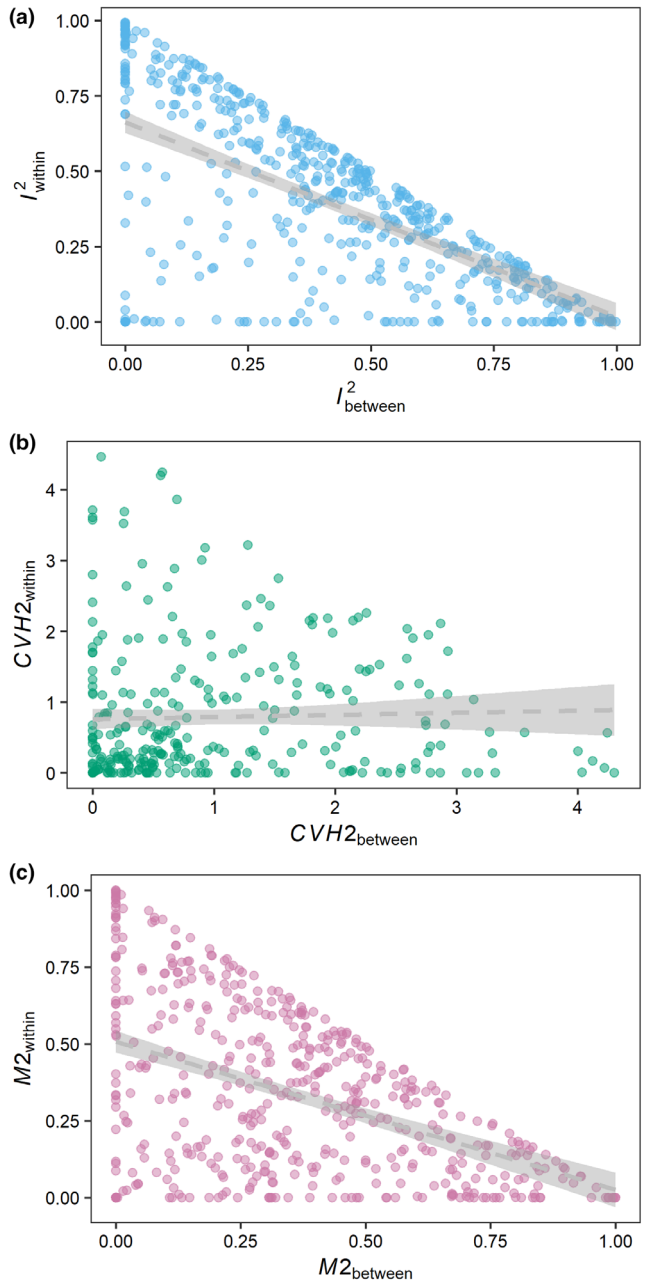

FIGURE 4 Comparison of stratified heterogeneity estimates across 512 meta-analyses for three heterogeneity metrics: (a) I^2, (b) coefficient of variation (*CVH2*) and (c) *M2*. Each point represents an estimate from an individual meta-analysis. Linear regression was applied to visualize trends (fitted lines), with shaded bands indicating the 95% confidence intervals. The correlation coefficients between between-study heterogeneity and within-study heterogeneity were −0.567, 95% CI = [−0.627, −0.500] for I^2; 0.482, 95% CI = [0.408, 0.549] for *CVH* 2; and −0.382, 95% CI = [−0.456, −0.303] for *M2*, respectively. Figures S6 and S7 present between- and within-study heterogeneity through alternative visualizations. Refer to Figure 3 for additional details.

$CVH2_{between}$ and $CVH2_{within}$. This similarity is expected as they can be mathematically transformed into one another using equations $MH2_{total} = CVH2_{total} / (1 + CVH2_{total})$ and $\text{logit}(MH2_{total}) = \log(CVH2_{total})$. The median values for both $CVH2_{total}$ and $MH2_{total}$ 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 $MH2_{total}$ also suggests that meta-analyses with high heterogeneity can possess a considerable likelihood of generalizability at the between-study level, given the low $MH2_{between}$ (as we pointed out above with I^2 family metrics). On average, there was a median $MH2_{between} = 0.3$ (SD is 43% of the overall mean effect) observed in 47% of the meta-analyses (242/512) with smaller $MH2_{between}$ values compared to $MH2_{within}$ values (Figure 4).

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

We found only moderate agreement between heterogeneity measured as I^2 and the newly proposed metrics ($CVH2_{total}$: $r_{spearman} = 0.319$, 95% CI = [0.237, 0.396], $MH2_{total}$: $r_{spearman} = 0.319$, 95% CI = [0.237, 0.396]; Figures 2b and 5a). In cases of meta-analyses with I^2 larger than 0.75 or smaller than 0.25 (identified as large and small heterogeneity by conventional benchmarks; Higgins et al., 2003), the disagreement between I^2 and *CVH2*, as well as I^2 and *MH2*, became even more pronounced (see Figures S8–S10 for additional results about inter-rater agreement test). In contrast, a near-perfect–though non-linear–relationship was observed between $CVH2_{total}$ and $MH2_{total}$ ($r_{spearman} = 1$, 95% CI = [0.999, 1]; Figure 5c). Therefore, cross-meta-analysis (meta-scientific) evidence suggests that I^2 as a measure of heterogeneity does not always agree with magnitude measures ($CVH2_{total}$ and $MH2_{total}$) for ecological and evolutionary data. We also found that out of the 512 meta-analyses featuring medium to large I^2_{total} values (>0.50 based on conventional guidelines), 80 had small $CVH2_{total}$ (Figure 5), indicating that more than 20% of the large I^2_{total} values were caused by small sampling errors rather than a larger amount of heterogeneity. These findings emphasize the importance of considering multiple metrics to obtain a holistic understanding of heterogeneity in meta-analyses (see Section 3.2).

3.2 | Heterogeneity interpretation benchmarks and a pluralistic framework

To support the interpretation of the newly proposed mean-standardized and variance–mean-standardized heterogeneity metrics, we derived empirical benchmarks based on their observed distributions across 512 ecological and evolutionary meta-analyses (illustrated in Figure 3). We tentatively classify heterogeneity as 'very small', 'small', 'moderate' or 'large' based on quartiles (specifically, the 0th to 25th, 25th to 50th, 50th to 75th and 75th to 100th percentiles). Table 1 presents these percentiles for variance-standardized heterogeneity (I^2 family metrics), mean-standardized (*CVH2* family metrics) and variance–mean-standardized (*M2* family metrics) heterogeneity. For example, for *M2*, the benchmarks for

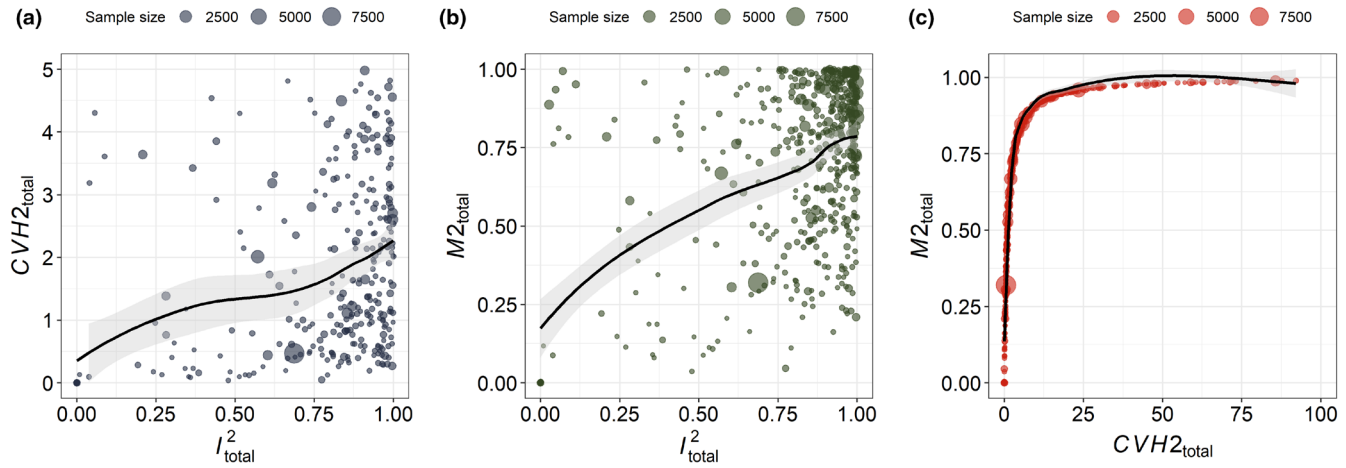

FIGURE 5 Comparison of heterogeneity measure estimates across 512 meta-analyses. A local polynomial regression was applied to illustrate trends (fitted lines), with shaded bands representing 95% confidence intervals. (a) I^2 exhibits moderate consistency with $CVH2_{total}$. (b) I^2 exhibits moderate consistency with $M2_{total}$. (c) $CVH2_{total}$ and $M2_{total}$ show a near-perfect-though non-linear-relationship. See Figure 3 for further details.

very small, small, moderate and large heterogeneity corresponded to values of 0 to 0.51, 0.51 to 0.78, 0.78 to 0.93 and 0.93 to 1, respectively (Table 1). Additionally, Table 2 offers a more fine-grained interpretation of these benchmarks for commonly used effect size measures, including SMD, lnRR and *Zr*. These empirically derived benchmarks are intended to provide general guidance for interpreting heterogeneity in ecological and evolutionary studies.

However, it is essential to recognize that these benchmarks should not replace contextual interpretation, which remains critical. Because effect size metrics differ in scale and statistical properties, heterogeneity estimates are inherently influenced by the metric used. Therefore, empirical benchmarks are most informative when applied to commonly used, standardized effect sizes (e.g. SMD, lnRR, Zr) and we discourage their application to less frequent metrics in our dataset (e.g. odds ratios or raw means), where coverage is too sparse for reliable guidance. When domain-specific knowledge is insufficient, these empirical benchmarks can serve as a starting point (rather than a substitute) for interpreting heterogeneity. In this spirit, we propose a pluralistic framework that encourages a comprehensive assessment of biological generalizability by jointly quantifying and contextualizing heterogeneity. Our key recommendations are as follows.

1. Adopt a multilevel meta-analytic framework: We strongly recommend modelling heterogeneity using multilevel meta-analysis (e.g. Equation 1) rather than a standard random effects model. The multilevel structure enables partitioning of heterogeneity across nested levels, and additional random effects (e.g. study ID, phylogeny, species) can be incorporated as needed. For instance, the phylogenetic multilevel model (Equation 12) can disentangle species-level sources of heterogeneity.
2. Quantify and stratify heterogeneity using complementary metrics: We encourage transparent reporting of all variance components, including average sampling error variance. From these components, multiple heterogeneity metrics can be computed, including I^2, *M* and *CVH* (the latter derivable from *M*), along with stratified versions. These measures provide complementary information, for example, I^2 quantifies the proportion of observed variance due to heterogeneity, whereas *M* and *CVH* place heterogeneity in the context of the mean. When there is not enough contextual information to guide the interpretation of heterogeneity, we encourage using the empirically derived benchmarks (Tables 1 and 2) as a starting point, particularly for commonly used effect size types such as SMD, lnRR and Zr. However, these benchmarks should not be used for less frequently used metrics in our dataset, such as 2×2 table-based measures (e.g. odds ratios) or non-standardized metrics like raw means, due to insufficient data coverage.
3. Use the R function (*het_interpret()*) to help obtain precise percentile ranges of heterogeneity estimates for a given meta-analysis based on empirical benchmarks (see online tutorial). The uncertainty (e.g. 95% CI) of each of the heterogeneity measures should be reported along with the point estimate, until a time when extensive simulation studies can provide a clear recommendation on which estimation methods provide the most reliable estimate of the uncertainty. Instead, we encourage researchers to conduct sensitivity analyses to understand the potential influence of heterogeneity (see 'Unresolved issue: quantifying the uncertainty around the point estimate of heterogeneity measure').
4. Check model parameter identifiability: When including multiple random effects, issues of parameter identifiability may arise, wherein unique variance estimates that maximize the likelihood function may not exist (see Section 2; Raue et al., 2009). Identifiability issues may arise in models with limited data or overlapping random effects. We recommend evaluating parameter identifiability (e.g. using profile likelihoods) before proceeding with the interpretation of heterogeneity estimates. If identifiability is uncertain, heterogeneity estimates should be interpreted with caution.

In conclusion, we advocate that ecologists and evolutionary biologists treat heterogeneity with the same importance as mean effect sizes when drawing biological inferences (Higgins et al., 2009). Our pluralistic approach provides the conceptual and practical tools to achieve this goal. We illustrate its implementation through two applied examples in our online tutorial (https://yefeng0920.github.io/heterogeneity_guide/).

3.3 | 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). Tackling such complexity can be achieved by embracing a flexible random effects structure within the multilevel meta-analytic framework (Nakagawa et al., 2023; Yang et al., 2022). As an example, we extend our models by introducing how heterogeneity can be decomposed into non-phylogenetic and phylogenetic species-level strata—a common set of random effects included in multilevel models (Cinar et al., 2022). Such an approach offers a unique opportunity for further disentangling heterogeneity and understanding generalisability.

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 as follows (Cinar et al., 2022; Nakagawa & Santos, 2012):

$$\mathrm{ES}_{[i]} = \mu + u_{\mathrm{species}[k]} + u_{\mathrm{phylogeny}[k]} + u_{\mathrm{between}[j]} + u_{\mathrm{within}[i]} + e_{[i]}, \tag{12}$$

where $u_{\mathrm{species}[k]}$ denotes the non-phylogenetic species random effect, which follows a normal distribution with mean zero and variance $\sigma^2_{\mathrm{species}}$; $u_{\mathrm{phylogeny}[k]}$ denotes the phylogenetic species random effect, which follows a normal distribution with mean zero and variance–covariance matrix $\sigma^2_{\mathrm{phylogeny}}\boldsymbol{A}$ (where $\sigma^2_{\mathrm{phylogeny}}$ is the phylogenetic species variance, and $\boldsymbol{A}$ is the 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 into the phylogenetic ($\sigma^2_{\mathrm{phylogeny}}$) and non-phylogenetic species level ($\sigma^2_{\mathrm{species}}$). Such stratification allows for the assessment of the heterogeneity within these strata, as illustrated in the empirical example below. Phylogenetic and non-phylogenetic species-level heterogeneity can be measured using $I^2_{\mathrm{phylogeny}}$ and I^2_{species}, respectively. These metrics are defined as follows:

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

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

where all notations are as previously defined. These expressions can also provide insights into the phylogenetic effect or signal, which reflects how shared evolutionary histories among species explain patterns of similarity (e.g. in trait values; Freckleton et al., 2002). One key parameter representing this concept is phylogenetic heritability, H^2 (Lynch, 1991; Nakagawa & Santos, 2012). While the definition of h^2 is not consistent in the literature (Pearse et al., 2023), a simple way to define H^2 is to exclude the sampling error variance $\bar{v}$ from equations (Nakagawa & Santos, 2012), resulting in the following expression:

$$H^2 = \frac{\sigma^2_{\mathrm{phylogeny}}}{\sigma^2_{\mathrm{phylogeny}} + \sigma^2_{\mathrm{species}} + \sigma^2_{\mathrm{between}} + \sigma^2_{\mathrm{within}}}. \tag{15}$$

Here, H^2 represents the proportion of variance attributed to phylogeny ($\sigma^2_{\mathrm{phylogeny}}$) relative to the total variance of the true effect sizes in the model. Therefore, when $H^2 = 0$, there is no phylogenetic effect or signal, whereas $H^2 = 1$ indicates that the effect sizes (or traits) among species are entirely determined by their phylogenetic relatedness. Another widely used parameter for assessing the phylogenetic signal is Pagel's λ (Cinar et al., 2022; Freckleton et al., 2002; Pagel, 1999), which is given by

$$\lambda = \frac{\sigma^2_{\mathrm{phylogeny}}}{\sigma^2_{\mathrm{phylogeny}} + \sigma^2_{\mathrm{species}}}. \tag{16}$$

Unlike h^2 (Equation 15), λ specifically reflects the proportion of phylogenetic variance ($\sigma^2_{\mathrm{phylogeny}}$) relative to the variance across species ($\sigma^2_{\mathrm{species}}$). Together, these parameters provide complementary perspectives on the role of phylogeny in shaping effect sizes or traits of interest (see Appendix S1 for the extended metrics for phylogenetic signal).

Following the same principle of Equations (7, 8, 10 and 11), we can derive the stratified version of mean-standardized and variance–mean-standardized heterogeneity measures. Mean-standardized heterogeneity metrics are given by

$$CVH2_{\mathrm{phylogeny}} = \frac{\sigma^2_{\mathrm{phylogeny}}}{\mu^2}, \tag{17}$$

$$CVH2_{\mathrm{species}} = \frac{\sigma^2_{\mathrm{species}}}{\mu^2}. \tag{18}$$

Variance–mean-standardized heterogeneity metrics are given by

$$M2_{\mathrm{phylogeny}} = \frac{\sigma^2_{\mathrm{phylogeny}}}{\sigma^2_{\mathrm{phypogeny}} + \sigma^2_{\mathrm{species}} + \sigma^2_{\mathrm{between}} + \sigma^2_{\mathrm{within}} + \mu^2}, \tag{19}$$

$$M2_{\mathrm{species}} = \frac{\sigma^2_{\mathrm{species}}}{\sigma^2_{\mathrm{phypogeny}} + \sigma^2_{\mathrm{species}} + \sigma^2_{\mathrm{between}} + \sigma^2_{\mathrm{within}} + \mu^2}. \tag{20}$$

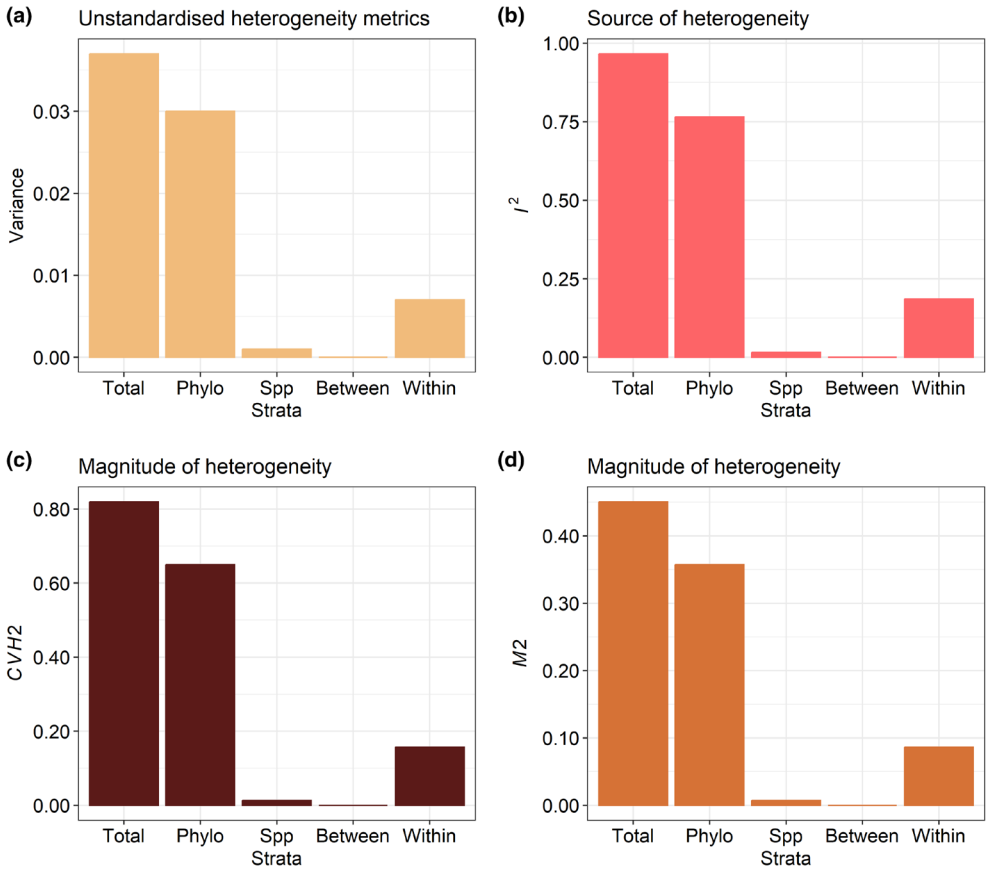

FIGURE 6 Heterogeneity quantification and stratification for multiple metrics. (a) Heterogeneity is quantified using the raw variance, (b) source measure, I^2, (c) magnitude measure, *CVH2* and (d) magnitude measure *M*, and stratified at phylogenetic (Phylo), non-phylogenetic (Spp), between-study (Between) and within-study (Within) levels. The source measure I^2 sometimes aligns well with the raw variance, as observed in this example (a, 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 variance-based effect size measures, such as the variation ratio and coefficient of variation ratio, as demonstrated in a real example at https://yefeng0920.github.io/heterogeneity_guide/.

One can also easily derive the variance—mean-standardized version of the phylogenetic signal index (see Appendix S1). To illustrate the insights gained through these extended measures, we present two case studies. The first case involves the phylogenetic meta-analysis originally conducted by Risely et al. (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 using standardized mean difference (SMD) as the effect size measure. The second case study involves the publicly available meta-analytic dataset about the impact of artificial light at night on the suppression of melatonin in wildlife (Yang, Liu, et al., 2024). While we reported our re-analysis of the first case study in the main text (Table S2), we reported the second one in the online tutorial due to limited space.

In our first case study, our re-analysis yielded two observations. Firstly, $I^2_{total} = 0.97$ exceeded the 85th percentile of the empirically derived heterogeneity distribution specific to SMD (Figure 6). 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 between the 25th and 50th percentiles of the empirically derived heterogeneity distribution specific to SMD via the R helper function *het_interpret()* ($CVH2_{total} = 1.3$ and $M2_{total} = 0.6$). This amount of heterogeneity can be tentatively interpreted as 'small to medium', compared to the heterogeneity of ecological and evolutionary meta-analyses using SMD as the effect size measure. This discrepancy was attributed to the tiny 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 contrary, we emphasize 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, $I^2_{total} = 0.97$ suggests the heterogeneity can explain most (97%) of the variability in the observed effect (only 3% is explained by the sampling error variance, or the heterogeneity is 32 times larger than that of sampling error variance).

Secondly, the effect of interest is highly likely to be generalizable and replicable at the between-study level when accounting for within-study variance. This conclusion is supported by the

stratification analysis, which reveals that between-study heterogeneity is extremely low, even though traditional benchmarks suggest substantial overall heterogeneity. An emerging approach in ecology and evolutionary biology, coordinated distributed experiments (Fraser et al., 2013), holds promise for controlling within-study variance by employing standardized and controlled protocols. Traditional meta-analytic practices risk overlooking such nuanced insights into heterogeneity and generalizability, potentially leading to erroneous conclusions. For instance, while random effects meta-analysis indicates high overall heterogeneity ($I^2_{total} = 0.96$; Figure 4; Table 1), stratification analysis shows that this heterogeneity is not driven by between-study differences. Instead, it is predominantly explained by phylogenetic effects ($I^2_{phylogeny} = 0.76$), which suggests that the mean effect is still generalizable across studies despite high total heterogeneity.

3.4 | Unresolved issue: Quantifying the uncertainty around the heterogeneity estimate

It is well recognized that the overall mean effect size for an outcome of interest should be reported with an uncertainty measure, such as a (95%) confidence interval, to indicate the precision of the estimate. In contrast, reporting confidence intervals for heterogeneity estimates is still uncommon. There are significant challenges to address before the routine construction of confidence intervals for heterogeneity becomes feasible. Several methods for constructing confidence intervals for unstandardized heterogeneity estimates (e.g. σ^2_{total}) have been proposed and tested (Viechtbauer, 2007). However, no established methods currently exist for standardized heterogeneity measures. Two simulation studies reveal that most approaches for constructing confidence intervals around unstandardized heterogeneity do not consistently achieve nominal coverage probabilities (Veroniki et al., 2016; Viechtbauer, 2007). For example, Wald-type and profile likelihood methods frequently yield coverage probabilities that deviate from the nominal level, while the Q-profile method can provide more accurate coverage under conditions that are closer to practical applications. However, the empirical performance of Q-profile remains untested in the multilevel modelling context (Equation 1).

For standardized heterogeneity measures, there are no established closed-form solutions or iterative procedures to construct confidence intervals. One straightforward approach involves using the confidence interval bounds for σ^2_{total} to calculate bounds for variance-standardized heterogeneity measures like I^2_{total}. Alternatively, the multivariate delta method could derive sampling variances and construct Wald-type confidence intervals for standardized heterogeneity measures, relying on the asymptotic normality of maximum likelihood and restricted maximum likelihood estimates. However, extensive simulation studies are necessary to assess the empirical performance of such intervals (e.g. power, Type I error rates) under conditions representative of multilevel models, which is beyond the scope of this paper. Bootstrapping methods also offer a possible solution for constructing confidence intervals for standardized heterogeneity measures. Yet, simulation studies indicate that both parametric and non-parametric bootstrap confidence intervals for unstandardized heterogeneity measures often exhibit suboptimal properties (Veroniki et al., 2016; Viechtbauer, 2007). Parametric bootstrapping assumes that parameter estimates (e.g. μ and σ^2_{total}) represent population parameters, disregarding their inherent uncertainty, while non-parametric bootstrapping fails to account for the multilevel structure of data.

Future research should focus on deriving uncertainty measures for the newly proposed standardized heterogeneity measures and conducting simulations to evaluate their performance in the context of a multilevel model. For now, researchers might conduct sensitivity analyses to address the limitations of ignoring uncertainty around heterogeneity estimates. For instance, a trace plot can illustrate the sensitivity of meta-analytic conclusions (e.g. the overall mean effect size estimate and its confidence interval) to changes in any heterogeneity measure (Röver et al., 2024). In such a plot, the x axis can represent different values of σ^2_{total} (or $CVH2_{total}$ and $M2_{total}$), while the y axis displays the corresponding overall mean effect size estimate (μ). Confidence intervals for σ^2_{total} based on Q-profile methods could facilitate these sensitivity analyses by suggesting a plausible range of values for consideration. Importantly, until reliable and validated methods for computing uncertainty ranges for standardized heterogeneity measures in the context of multilevel meta-analysis, reporting 95% confidence intervals or other uncertainty estimates for these metrics should be interpreted with caution.

AUTHOR CONTRIBUTIONS

Yefeng Yang: Conceptualization; data curation; formal analysis; investigation; methodology; software; visualization; writing—original draft; writing—review and editing. Daniel W. A. Noble: Software; visualization; writing—review and editing. Rebecca Spake: Writing—review and editing. Alistair M. Senior: Writing—review and editing. Malgorzata Lagisz: Visualization; writing—review and editing; funding acquisition. Shinichi Nakagawa: Conceptualization; investigation; methodology; software; validation; writing—review and editing; funding acquisition; supervision. All authors approved the final manuscript.

ACKNOWLEDGEMENTS

YY, SN and ML were funded by the Australian Research Council Discovery Grant (DP210100812 & DP230101248). DWAN was supported by an ARC Future Fellowship (FT220100276). AMS was supported by an ARC Future Fellowship (FT230100240). SN acknowledges support from a Canada Excellence Research Chair (CERC-2022-00074). Open access publishing facilitated by University of New South Wales, as part of the Wiley - University of New South Wales agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

All authors declare no competing interests.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70155.

DATA AVAILABILITY STATEMENT

Raw data and analytical script to reproduce results and figures presented in the manuscript is archived at Zenodo (https://doi.org/10.5281/zenodo.15718008; Yang, 2025) and GitHub repository (https://github.com/Yefeng0920/heterogeneity_benchmark). A webpage showing the use of the proposed method can be accessed via https://yefeng0920.github.io/heterogeneity_guide/.

ORCID

Yefeng Yang https://orcid.org/0000-0002-8610-4016
Daniel W. A. Noble https://orcid.org/0000-0001-9460-8743
Rebecca Spake https://orcid.org/0000-0003-4671-2225
Alistair M. Senior https://orcid.org/0000-0001-9805-7280
Malgorzata Lagisz https://orcid.org/0000-0002-3993-6127
Shinichi Nakagawa https://orcid.org/0000-0002-7765-5182

REFERENCES

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 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*, 211–229.

Chik, H. Y. J., Sparks, A. M., Schroeder, J., & Dugdale, H. L. (2022). A meta-analysis on the heritability of vertebrate telomere length. *Journal of Evolutionary Biology*, *35*, 1283–1295.

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.

Fraser, L. H., Henry, H. A., Carlyle, C. N., White, S. R., Beierkuhnlein, C., Cahill, J. F., Jr., Casper, B. B., Cleland, E., Collins, S. L., & Dukes, J. S. (2013). Coordinated distributed experiments: An emerging tool for testing global hypotheses in ecology and environmental science. *Frontiers in Ecology and the Environment*, *11*, 147–155.

Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic analysis and comparative data: A test and review of evidence. *The American Naturalist*, *160*, 712–726.

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.

Heidenreich, N.-B., Schindler, A., & Sperlich, S. (2013). Bandwidth selection for kernel density estimation: A review of fully automatic selectors. *AStA Advances in Statistical Analysis*, *97*, 403–433.

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.

Kvålseth, T. O. (2017). Coefficient of variation: The second-order alternative. *Journal of Applied Statistics*, *44*, 402–415.

Lobry, J. R., Bel-Venner, M. C., Bogdziewicz, M., Hacket-Pain, A., & Venner, S. (2023). The CV is dead, long live the CV! *Methods in Ecology and Evolution*, *14*, 2780–2786.

Lynch, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution*, *45*, 1065–1080.

Monaghan, P., Eisenberg, D. T., Harrington, L., & Nussey, D. (2018). Understanding diversity in telomere dynamics. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *373*, 20160435.

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.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*, 877–884.

Pearse, W. D., Davies, T. J., & Wolkovich, E. (2023). How to define, use, and interpret Pagel's λ (lambda) in ecology and evolution. *bioRxiv*, 2023.2010.2010.561651.

Pick, J. L., Kasper, C., Allegue, H., Dingemanse, N. J., Dochtermann, N. A., Laskowski, K. L., Lima, M. R., Schielzeth, H., Westneat, D. F., & Wright, J. (2023). Describing posterior distributions of variance components: Problems and the use of null distributions to aid interpretation. *Methods in Ecology and Evolution*, *14*, 2557–2574.

R Core Team. (2020). *R: A language and environment for statistical computing*. Foundation for Statistical Computing.

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.

Remot, F., Ronget, V., Froy, H., Rey, B., Gaillard, J. M., Nussey, D. H., & Lemaitre, J. F. (2022). Decline in telomere length with increasing age across nonhuman vertebrates: A meta-analysis. *Molecular Ecology*, *31*, 5917–5932.

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.

Röver, C., Rindskopf, D., & Friede, T. (2024). How trace plots help interpret meta-analysis results. *Research Synthesis Methods*, *15*, 413–429.

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, *8*, 1–9.

Salmón, P., & Burraco, P. (2022). Telomeres and anthropogenic disturbances in wildlife: A systematic review and meta-analysis. *Molecular Ecology*, *31*, 6018–6039.

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*, *6*, 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.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*, 55–79.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*, 37–52.

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. (2025). Heterogeneity interpretation (v1.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.15718008

Yang, Y., Lagisz, M., Williams, C., Noble, D. W., Pan, J., & Nakagawa, S. (2024). Robust point and variance estimation for meta-analyses with selective reporting and dependent effect sizes. *Methods in Ecology and Evolution*, *15*, 1593–1610.

Yang, Y., Liu, Q., Pan, C., Chen, J., Xu, B., Liu, K., Pan, J., Lagisz, M., & Nakagawa, S. (2024). Species sensitivities to artificial light at night: A phylogenetically controlled multilevel meta-analysis on melatonin suppression. *Ecology Letters*, *27*, e14387.

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*, *146*, 105016.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Table S1. Descriptive summary of datasets excluded due to model convergence problems.

Table S2. Results of heterogeneity quantification and stratification based on multiple measures.

Table S3. Rule-of-thumb and empirically derived benchmarks for the interpretation of the full set of standardised heterogeneity metrics.

Figure S1. Illustration of coefficient of variation (*CV*).

Figure S2. Histogram of heterogeneity estimates derived from 512 meta-analyses was systematically assessed using pluralistic measures and stratified across different strata.

Figure S3. The untruncated distribution of heterogeneity estimates of $CVH2$ derived from 512 meta-analyses.

Figure S4. The distribution of estimates of total variance in effect size (σ^2_{total}) derived from 512 meta-analyses.

Figure S5. The distribution of estimates of typical sampling error variance in effect size ($\bar{v}$) derived from 512 meta-analyses.

Figure S6. Paired comparison of stratified heterogeneity estimates across 512 meta-analyses for three heterogeneity metrics: (A) I^2, (B) coefficient of variation ($CVH2$), and (C) $M2$.

Figure S7. Paired comparison of stratified heterogeneity estimates derived 512 meta-analyses for untruncated $CVH2$.

Figure S8. The agreement chart for a 3×3 contingency table (confusion matrix) that assesses the congruence between I^2 and $CVH2$ in interpreting heterogeneity magnitude.

Figure S9. The agreement chart providing a visual assessment of the congruence between I^2 and $MH2$ in interpreting heterogeneity magnitude.

Figure S10. The agreement chart provides a visual assessment of the congruence between $M2$ and $CVH2$ in interpreting heterogeneity magnitude.

Appendix S1. Supplementary methodologies including technical explanation of the principle of decomposing meta-analytic heterogeneity and their extended metrics.

How to cite this article: Yang, Y., Noble, D. W. A., Spake, R., Senior, A. M., Lagisz, M., & Nakagawa, S. (2025). A pluralistic framework for measuring, interpreting and decomposing heterogeneity in meta-analysis. *Methods in Ecology and Evolution*, *16*, 2710–2725. https://doi.org/10.1111/2041-210x.70155