

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

3

4 Yefeng Yang<sup>1</sup>, Daniel W. A. Noble<sup>2</sup>, Rebecca Spake<sup>3</sup>, Alistair M. Senior<sup>4</sup>, Malgorzata  
5 Lagisz<sup>1,#</sup>, Shinichi Nakagawa<sup>1,5,#</sup>

6

7 <sup>1</sup>Evolution & Ecology Research Centre and School of Biological, Earth and Environmental  
8 Sciences, University of New South Wales, Sydney, NSW 2052, Australia

9 <sup>2</sup>Division of Ecology and Evolution, Research School of Biology, The Australian National  
10 University, Canberra, ACT 2600, Australia

11 <sup>3</sup>Ecology and Evolutionary Biology Research Division, School of Biological Sciences,  
12 University of Reading, RG6 6EX, Reading, UK

13 <sup>4</sup>Charles Perkins Centre, Sydney Precision Data Science Centre, School of Life and  
14 Environmental Sciences and School of Mathematics and Statistics, The University of Sydney,  
15 Sydney, NSW 2006, Australia

16 <sup>5</sup>Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology  
17 Graduate University, Onna, 904-0495, Japan

18

19 Correspondence

20 E-mail: [s.nakagawa@unsw.edu.au](mailto:s.nakagawa@unsw.edu.au) (SN)

21 # Equal senior author

22

23 **ORCID**

24 Yefeng Yang 0000-0002-8610-4016

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

- 26 Rebecca Spake 0000-0003-4671-2225
- 27 Alistair M. Senior 0000-0001-9805-7280
- 28 Malgorzata Lagisz 0000-0002-3993-6127
- 29 Shinichi Nakagawa 0000-0002-7765-5182
- 30

31 **Abstract**

32 Uncovering general rules enhances the predictive capabilities in ecology and evolution. Meta-  
33 analytic approaches play a critical role in this endeavour, examining the extent to which  
34 phenomena can be replicated, generalized, and transferred. However, ecologists and  
35 evolutionary biologists have largely overlooked the role of meta-analytic heterogeneity in  
36 informing generality. To reform this situation, we introduce a pluralistic approach aimed at  
37 quantifying and stratifying various heterogeneity metrics, such as  $I^2$ ,  $CV$ ,  $M$ , and predictive  
38 distribution. These metrics offer complementary information, revealing the source,  
39 magnitude, and visual representation of heterogeneity. Our analysis of 512 meta-analyses  
40 demonstrates that heterogeneity is, on average, ten times larger than statistical noise,  
41 contributing to 91% of the observed variance (median  $I^2 = 91\%$ ). This amount of  
42 heterogeneity is nearly twice the size of the meta-analytic mean effect (median  $CV = 1.8$ ,  $M =$   
43  $0.6$ ), indicating substantial total heterogeneity in ecology and evolution. Surprisingly, in half  
44 of the cases, focal effects could generalize across studies even with high total heterogeneity  
45 by controlling for within-study variation. Our synthesis also visualises empirical distributions  
46 of various heterogeneity metrics, potentially serving as new benchmarks for informed  
47 interpretation. Our proposed pluralistic approach will accelerate the future quest for general  
48 rules via meta-analyses.

49

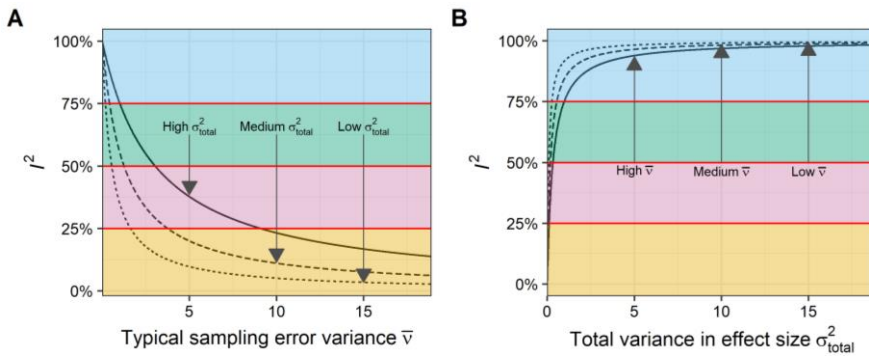
50 **Main**

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

62  
63 Meta-analyses play a crucial role in evaluating the generality of patterns <sup>3</sup>. Firstly, they  
64 quantitatively estimate the population mean effect across studies <sup>5-7</sup>, characterising the central  
65 tendency of a focal effect. Secondly, they can identify effect modifiers or moderators  
66 contributing to context dependence <sup>5</sup> and provide tailored estimates for target contexts <sup>4</sup>.  
67 Third, meta-analyses can quantify variability in study outcome, the “heterogeneity” among  
68 effect sizes. Without quantifying heterogeneity, it is difficult to interpret both the overall  
69 trends and context-specific effects <sup>8</sup>. Heterogeneity can help to indicate the degree of  
70 inconsistency, or context dependence, of study findings, with high heterogeneity signalling a  
71 need to investigate the drivers of the variation. Lower heterogeneity can indicate high  
72 generality. Specifically, the mean effect size is highly transferable across the contexts  
73 characterised by the study pool without the need to consider effect modifiers <sup>2</sup>. Until now, the

74 significance of heterogeneity in informing generality has been largely overlooked. Indeed,  
75 surveys have revealed that heterogeneity statistics are not routinely reported<sup>7-9</sup>.

76



77

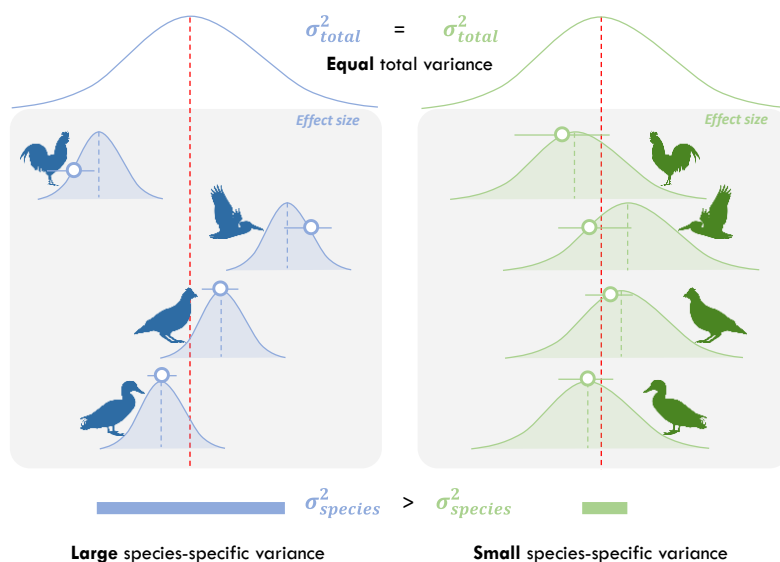
78 **Fig. 1:**

79 The interpretation of total  $I^2$  can be ambiguous and can lead to incorrect conclusions about the  
80 magnitude of heterogeneity. (A) A large estimated total  $I^2$  value could be due to small sampling error  
81 variances  $\bar{v}$  (i.e., low statistical noise). (B) On the other hand, a large total  $I^2$  value could also result  
82 from a large true heterogeneity. Values of  $\sigma_{total}^2$  and  $\bar{v}$  were derived from their empirical distributions  
83 based on 512 meta-analyses (see Figs. S1 and S2). Total  $I^2$  values were calculated using Equations 2  
84 and 3. High, medium, and low  $\sigma_{total}^2$  (and  $\bar{v}$ ) denote the 25%, 50%, and 75% percentiles of their  
85 empirical distributions (Table 1). Three horizontal lines denote the conventional thresholds for the use  
86 of  $I^2$  to interpret the magnitude of heterogeneity<sup>10</sup>.

87

88 Currently, measuring and interpreting meta-analytic heterogeneity faces two major  
89 limitations. First, no single heterogeneity metric provides a holistic interpretation of  
90 generality<sup>11</sup>. Currently, the  $I^2$  statistic is a popular metric that quantifies the proportion of  
91 variance due to differences between effect sizes rather than by statistical noise (i.e., sampling  
92 variance)<sup>12,13</sup>. The biological interpretation of  $I^2$ , however, is ambiguous<sup>14</sup> because a small  
93 absolute heterogeneity can lead to a high  $I^2$  due to small statistical noise (see Fig. 1)<sup>12,14,15</sup>. In

94 addition,  $I^2$  is a point estimate and cannot reflect the whole distribution of context-specific  
 95 effects <sup>16</sup>. Second, meta-analyses typically focus on estimating total heterogeneity only <sup>5</sup>,  
 96 despite the hierarchical nature of real biological data structures <sup>6,9</sup>. Explicitly decomposing  
 97 effect size heterogeneity across hierarchical levels (i.e., stratification) enables a more nuanced  
 98 assessment of generality, and helps in identifying contextual factors <sup>5</sup> that drive context  
 99 dependence. For example, in a multi-taxon meta-analysis, if stratification of studies by  
 100 species yields low heterogeneity at the taxon level, the focal effect still can be generalizable  
 101 across taxon (in terms of accounting for within-taxon variation; Fig. 2). This is so, even if the  
 102 total heterogeneity remains high <sup>8</sup>.



103

104 **Fig. 2:**

105 A cross-taxa meta-analysis with a high total variance can have a small amount of species-specific  
 106 heterogeneity. The focal effect is still possible to be generalizable at the species level. The circles  
 107 represent the replication of species-specific effect. The red dashed lines denote the meta-analytic  
 108 mean effects. See a real example in **Modelling additional source heterogeneity**.

109

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

## 123 **Discerning biological generality**

### 124 **Heterogeneity in multilevel meta-analytic modelling framework**

125 Data used in meta-analyses often exhibit a complex hierarchical structure<sup>5,19</sup>, with study  
126 identity serving as a typical clustering variable, forming two strata (or more). Ecological and  
127 evolutionary meta-analyses typically report around eight effect sizes per study<sup>20</sup>. However,  
128 Traditional random-effects meta-analytic approaches do not account for heterogeneity driven  
129 by such data stratification<sup>6,7,9</sup>, and multi-level meta-analysis is required to model  
130 heterogeneity at different strata or multi-levels in a meta-analysis (see **Methods**).

131

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

$$135 \quad ES_{[i]} = \mu + u_{between[j]} + u_{within[i]} + e_{[i]}, (1)$$

136 The typical assumptions for Equation 1 is as follows: (i) between-study-level random effect  
137  $u_{b[j]}$  follows a normal distribution with mean zero and variance  $\sigma_{between}^2$ :  $u_{between[j]} \sim$   
138  $\mathcal{N}(0, \sigma_{between}^2)$ , (ii) within-study-level random effect  $u_{within[i]}$  follows a normal distribution  
139 with mean zero and variance  $\sigma_{within}^2$ :  $u_{within[i]} \sim \mathcal{N}(0, \sigma_{within}^2)$ , and (iii) sampling error  $e_{[i]}$   
140 follows a normal distribution with mean zero and variance in effects defined by the sampling  
141 variance ( $v_{[i]}$ ) associated with each effect size,  $i$ , such that  $e_{[i]} \sim \mathcal{N}(0, v_{[i]})$ . The assumption  
142 of homogeneous variances for the random effects can be relaxed to allow for  
143 heteroscedasticity<sup>21</sup>. Similarly, the assumption of independent sampling errors ( $e_{[i]}$ ) can be  
144 relaxed to allow for sampling error covariance  $v_{[i]}$ <sup>7</sup>. In the following sections, we will  
145 elaborate on how to stratify heterogeneity information using Equation 1.

146



147 **Unstandardised heterogeneity metrics**

148 Cochran's  $Q$  is a widely used metric for assessing heterogeneity in meta-analyses<sup>22</sup>. It serves  
149 as a test statistic to determine whether the true effects are homogeneous or not, informing a  
150 binary decision as to whether the effect sizes come from a common underlying population, or  
151 not (i.e., is heterogeneity 'non-zero'?). In contrast, the variance of true effects ( $\sigma_{total}^2 =$   
152  $\sigma_{between}^2 + \sigma_{within}^2$ ) provides a direct measure of absolute heterogeneity. Equation 1 offers a  
153 general way to partition the variance of the observed effects into sampling error variance, and  
154 that of true effects at different strata, such as between-study ( $\sigma_{between}^2$ ) and within-study  
155 strata ( $\sigma_{within}^2$ ). By considering additional strata, such as variation in effects among species or  
156 geographical locations, the total variance in true effects ( $\sigma_{total}^2$ ) can be further decomposed to  
157 assess generality at these specific strata (See **Model additional source heterogeneity**). For  
158 example, high variation among studies implies lack of generality from one study to another  
159 while low variation among species implies effects are similar, on average, across species.  
160 Nonetheless, relying solely on such absolute variance may not provide practical intuition  
161 regarding the magnitude of heterogeneity. For example, in a meta-analysis with  $\sigma_{total}^2 = 1$ , it  
162 is unclear whether this amount of variance is large and meaningful because absolute variance  
163 is not unit-free and not comparable across effect size measure used.

164

165 **Variance-standardised heterogeneity metrics**

166 The heterogeneity index,  $I^2$  has emerged as the most popular metric as it provides a  
167 standardized measure of heterogeneity that accounts for the scale dependence (i.e., unit-free)  
168 <sup>10</sup>.  $I^2$  is a variance-scaled heterogeneity metric that measures the proportion of total variance  
169 beyond statistical noise<sup>13</sup>. The total  $I^2$  can be computed by dividing the variance in the true  
170 effects ( $\sigma_{total}^2$ ) by the variance in the observed effects ( $\text{Var}[ES_{[i]}]$ ), as follows:

171 
$$I_{total}^2 = \frac{\sigma_{total}^2}{\text{Var}[ES_{[i]}]} = \frac{\sigma_{total}^2}{\sigma_{total}^2 + \bar{v}}, (2)$$

172 where  $\bar{v}$  represents the “typical” sampling error variance.  $\bar{v}$  can be computed using different  
 173 estimators<sup>23,24</sup>, with the common one being<sup>13</sup>:

174 
$$\bar{v} = \frac{(k-1) \sum_{i=1}^k 1/v_{[i]}}{(\sum_{i=1}^k 1/v_{[i]})^2 - \sum_{i=1}^k 1/v_{[i]}^2}, (3)$$

175 Within the multilevel modelling framework, the total  $I^2$  can be stratified at different strata  
 176<sup>5,24</sup>, for example, by estimating  $I^2$  at between-study ( $I_{between}^2$ ) and within-study ( $I_{within}^2$ )  
 177 levels:

178 
$$I_{between}^2 = \frac{\sigma_{between}^2}{\text{Var}[ES_{[i]}]} = \frac{\sigma_{between}^2}{\sigma_{total}^2 + \bar{v}}, (4)$$

179 
$$I_{within}^2 = \frac{\sigma_{within}^2}{\text{Var}[ES_{[i]}]} = \frac{\sigma_{within}^2}{\sigma_{total}^2 + \bar{v}}, (5)$$

180 However, as mentioned earlier, large  $I^2$  values do not necessarily imply a practically relevant  
 181 amount of heterogeneity (see Fig. 1; also see a case study in **Model additional source of**  
 182 **heterogeneity**). Statistical noise can sometimes inflate  $I^2$  values, which is a common  
 183 occurrence in ecology and evolutionary meta-analyses (see **Empirical patterns of**  
 184 **heterogeneity in ecology and evolution**). Stratified  $I^2$  metrics range from 0 to 100% (but  
 185 together sum to 100%), providing a clearer intuition of the source of heterogeneity and aiding  
 186 in assessing the drivers of context dependence at different strata. For example, a  $I_{within}^2$  of  
 187 90% means within-study variation can account for 90% of heterogeneity, therefore, indicating  
 188 that within-study level predictors are more likely to drive context dependence.  $I^2$  and its  
 189 stratified variants can also be transformed into the ratio of the variance of true effect to  
 190 typical sampling error variance ( $\frac{\sigma^2}{\bar{v}} = \frac{I^2}{(1-I^2)}$  or  $\log\left(\frac{\sigma^2}{\bar{v}}\right) = \text{logit}(I^2)$ ), which represents  
 191 heterogeneity as a proportion of the statistical noise (sampling error variance).

192

193 **Mean-standardised heterogeneity metrics**

194 Evolutionary biologists and behavioural ecologists are familiar with the variance-scaled  
195 metrics such as heritability ( $h^2$ ) and repeatability ( $R$ ), which are statistically comparable to  
196 the heterogeneity index,  $I^2$ . Although less commonly used, there also exists the mean-scaled  
197 counterparts, such as evolvability or the coefficient of variation ( $CV$ ) for additive genetic  
198 variance ( $CV_A$ ) and  $CV$  for between-individual variance ( $CV_B$ )<sup>25</sup>. In a similar manner, there  
199 exists a mean-scaled heterogeneity metric that can provide a standardized measure of  
200 heterogeneity, denoted as  $CV_{total}$ , that compares the standard deviation  $\sigma_t$  to the magnitude  
201 of its population mean ( $\mu$ )<sup>23</sup>:

202 
$$CV_{total} = \frac{\sigma_{total}}{|\mu|}, (6)$$

203  $CV_t$  expresses the total heterogeneity as a proportion of the meta-analytic mean effect (or as a  
204 percentage of change in the meta-analytic mean effect when multiplied by 100). To provide a  
205 more precise quantification of heterogeneity at different strata, we propose stratified versions  
206 of  $CV_t$ . Under the simplest multilevel model framework (Equation 1), we propose estimating  
207 between-study,  $CV_b$ , and within-study,  $CV_w$ , as follows:

208 
$$CV_{between} = \frac{\sigma_{between}}{|\mu|}, (7)$$

209 
$$CV_{within} = \frac{\sigma_{within}}{|\mu|}, (8)$$

210 Notably, these mean-scaled variance metrics have the limitation of becoming arbitrarily large  
211 as the magnitude of meta-analytic mean effect  $|\mu|$  approaches zero<sup>26</sup>. It is this limitation that  
212 has probably prevented the widespread adoption of the mean-scaled variance in the field of  
213 evolutionary quantitative genetic and animal personality research<sup>25,27</sup>.

214

215 **Variance-mean-standardised heterogeneity metrics**

216 To remedy the problems of  $I^2$  and  $CV_{total}$  as illustrated above, there is a more robust measure  
217 of heterogeneity  $M_{total}$  that combines the strengths of mean-scaled and variance-scaled  
218 metrics <sup>11</sup>:

$$219 \quad M_{total} = \frac{\sigma_{between} + \sigma_{within}}{\sigma_{between} + \sigma_{within} + |\mu|}, (9)$$

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

$$223 \quad M_{between} = \frac{\sigma_{between}}{\sigma_{between} + \sigma_{within} + |\mu|}, (10)$$

$$224 \quad M_{within} = \frac{\sigma_{within}}{\sigma_{between} + \sigma_{within} + |\mu|}, (11)$$

225  $M_t$  and its stratified variants are still standardised measures that quantify the size of  
226 heterogeneity relative to the magnitude of meta-analytic mean effect, providing intuitive  
227 interpretation. For example,  $\sigma_{total} = 0$  leads to  $M_{total} = 0$ , indicating the population mean  
228 effect is fully generalisable, and replicable across different contexts (see a case study in  
229 **Model additional source of heterogeneity**). On the other hand,  $M_{total}$  and its stratified  
230 variants are truncated at one, which overcomes the issue of  $CV_{total}$  when the magnitude of  
231 meta-analytic mean effect  $|\mu|$  approaches zero. Note that there is another mean- and variance-  
232 scaled metric,  $M_{total}^2$ , where  $\sigma_{total}$  and  $|\mu|$  are replaced by their squared values (**Methods**).  
233  $CV_{total}$ ,  $M_{total}$  and  $M_{total}^2$  can be all be easily stratified using multilevel meta-analytic  
234 models (**Model additional source of heterogeneity**).

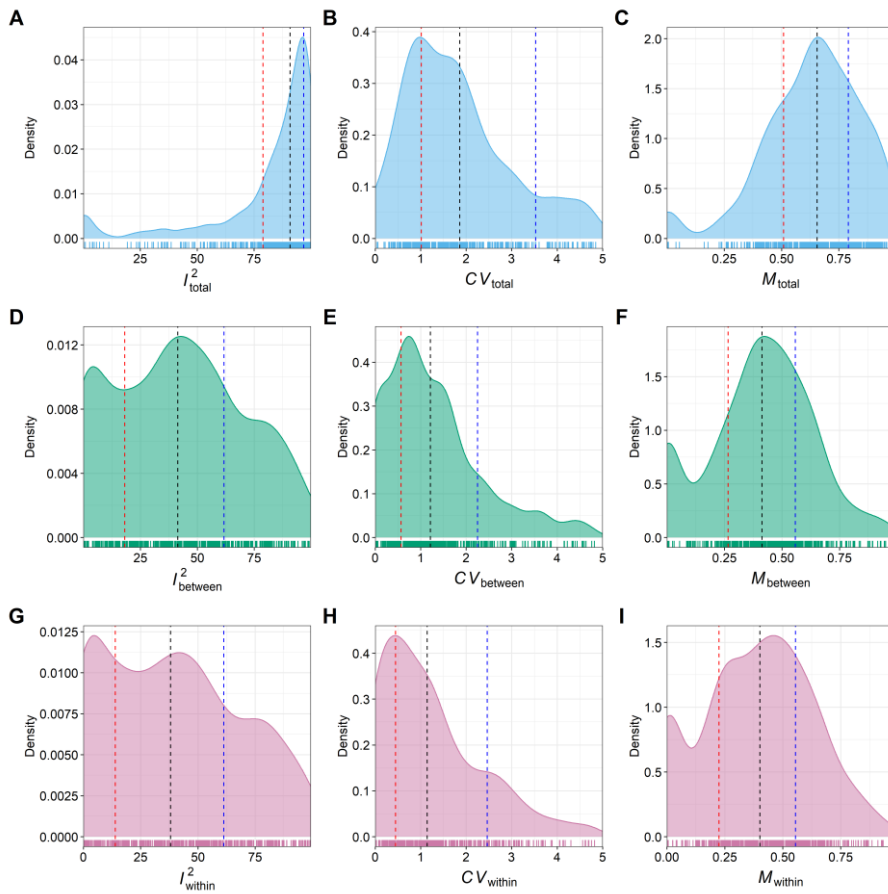
235

### 236 **Empirical patterns of heterogeneity in ecology and evolution**

237 To evaluate empirical patterns in heterogeneity among meta-analytic studies in ecology and  
238 evolution, we applied multilevel meta-analytic models (Equation 1) to 512 published meta-

239 analyses<sup>18,28</sup>. For each meta-analysis, we quantified and stratified heterogeneity using  $I_{total}^2$ ,  
240  $CV_{total}$ ,  $M_{total}$ . For  $I_{total}^2$ , the 25th, 50th, and 75th percentiles corresponded to 79%, 91%,  
241 and 97%  $I_{total}^2$ , respectively (Fig. 3), rather than conventional thresholds for interpreting  $I^2$ ,  
242 which typically categorize heterogeneity as small, moderate, or high at 25%, 50%, and 75%  
243  $I_{total}^2$ , respectively<sup>10</sup>. This also means, on average, variation in true effect sizes  $\sigma^2$  was ten  
244 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  
245 empirical distributions of  $\sigma^2$  and  $\bar{v}$ ) and 91% of them can be attributed to the ‘true’ biological  
246 or methodological differences in research contexts, and thus are theoretically explainable  
247 using appropriate predictors.

248  
249 While  $I_{total}^2$  displayed a left-skewed and single-modal distribution, its stratified counterparts,  
250  $I_{between}^2$  and  $I_{within}^2$ , demonstrated a right-skewed distribution with multi-modal patterns.  
251 There was no consistent trend suggesting one type of stratified heterogeneity consistently  
252 outweighed the other across the 512 meta-analyses (Fig. 3). Intriguingly, 47% (242 out of  
253 512) of the meta-analyses exhibited smaller between-study level heterogeneity than within-  
254 study level heterogeneity ( $I_{between}^2 < I_{within}^2$ ; Fig. 4). Within this subset of meta-analyses, the  
255 median values for  $I_{total}^2$ ,  $I_{between}^2$  and  $I_{within}^2$  were 95%, 21%, and 63%, respectively. It  
256 highlights a key finding often overlooked by traditional heterogeneity quantification  
257 practices: findings from many meta-analyses with high total heterogeneity can still be  
258 generalized at the between-study study level. Such generalization is achievable when  
259 replication is defined as the testing of the null hypothesis at the between-study level, and  
260 when within-study methodological and biological variations can be adequately accounted for  
261 (i.e., within-lab heterogenization<sup>29</sup>) because some meta-analyses have relatively low  
262 heterogeneity at the between-study study level.



263

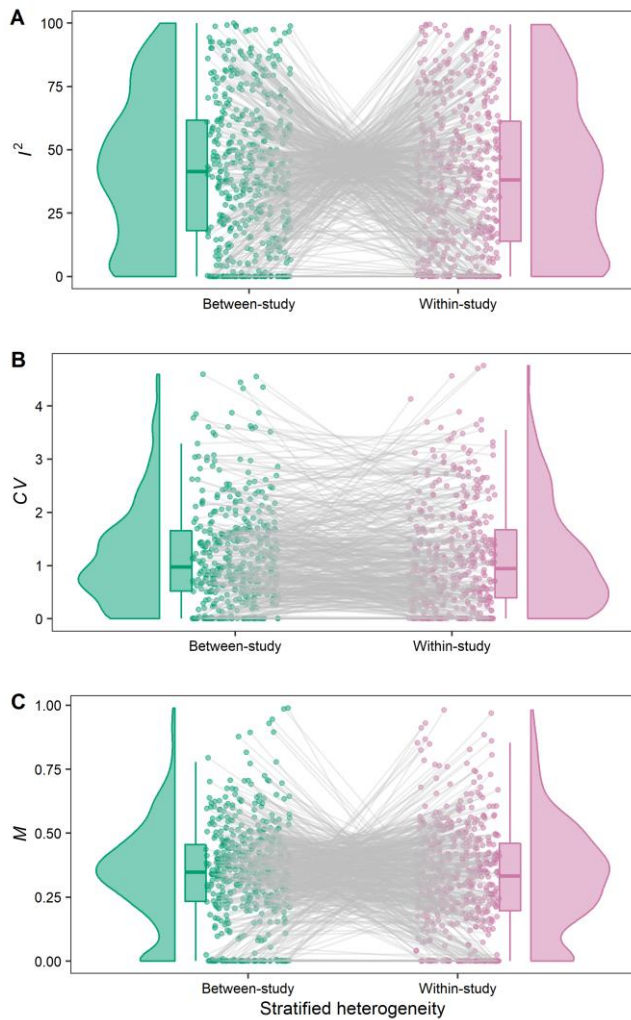
264 **Fig. 3:**

265 The distribution of heterogeneity estimates derived from 512 meta-analyses was systematically  
 266 assessed using pluralistic measures and stratified across different strata. Total heterogeneity measures  
 267 (A – C):  $I^2_{total}$ ,  $CV_{total}$  and  $M_{total}$ . Between-study heterogeneity measures (D – E):  $I^2_{between}$ ,  
 268  $CV_{between}$  and  $M_{between}$ . Within-study heterogeneity measures (G – I):  $I^2_{within}$ ,  $CV_{within}$  and  
 269  $M_{within}$ . Three dashed lines correspond to the 25th, 50th, and 75th percentiles, respectively. In panels  
 270 B, E, and H, the  $CV$  was truncated at five for figure clarity, as very large  $CV$  values can be challenging  
 271 to interpret when the meta-analytic mean effect is small. For example, the maximum  $CV$  observed in

272 the 512 meta-analyses was 106, which was inflated by a small meta-analytic mean effect of 0.03. For  
273 the figures without truncation, please refer to Figure S3.

274

275 When the  $CV_{total}$  metric was used to quantify heterogeneity, the calculated 25th, 50th, and  
276 75th percentiles of  $CV_{total}$  values were 1.0, 1.8, and 3.5, respectively (Fig. 3). This means  
277 that the standard deviation (in this case, heterogeneity) was, on average, nearly twice that of  
278 the meta-analytic mean effect. The distributions of both  $CV_{total}$  and its stratified versions,  
279  $CV_{between}$ , and  $CV_{within}$ , displayed a right-skewed pattern with a single-mode. In contrast, the  
280 distribution of  $M_t$  exhibited a more symmetrical pattern, with the 25th, 50th, and 75th  
281 percentiles of  $M_{total}$  values being 0.5, 0.6, and 0.8, respectively (Fig. 3), albeit with a minor  
282 peak around zero. Notably, stratification analysis revealed that  $M_{between}$  and  $M_{within}$  had  
283 patterns similar to those observed for  $CV_{between}$  and  $CV_{within}$ . This similarity is expected as  
284 they can be mathematically transformed into one another using equations  $M_{total} =$   
285  $CV_{total}/(1 + CV_{total})$  and  $logit(M_{total}) = \log(CV_{total})$ . The median values for both  
286  $CV_{total}$  and  $M_{total}$  across the 512 meta-analyses signify a high amount of heterogeneity,  
287 thereby warranting a thorough exploration into the drivers influencing such context  
288 dependence. However, stratification of  $M_{total}$  also suggests that meta-analyses with high  
289 heterogeneity can possess a considerable likelihood of generality at the between-study level,  
290 given low  $M_{between}$  (as we pointed out above with  $I^2$ ). On average, there was a median  
291  $M_{between} = 0.3$  (SD is 41% of the meta-analytic mean effect) observed in 47% of the meta-  
292 analyses (242/512) with smaller  $M_{between}$  values compared to  $M_{within}$  values (Fig. 4).



293

294 **Fig. 4:**

295 Paired comparison of stratified heterogeneity estimates derived 512 meta-analyses for three

296 heterogeneity metrics (A)  $I^2$ , (B) coefficient of variation,  $CV$  and (C)  $M$ . Heterogeneity was stratified

297 at both 'between-study' and 'within-study' levels (x-axes). Each point represents an estimate from

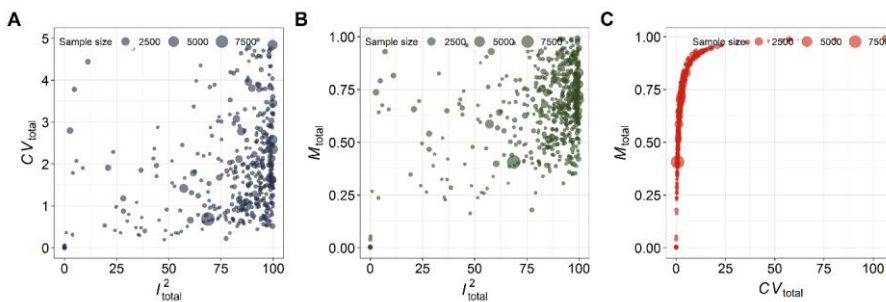
298 each meta-analysis. For panel B,  $CV$  has been truncated at five for figure clarity. For the full figures

299 without truncation, please refer to Figure S4. For other details see Fig. 3.

300



301 We found only moderate agreement between heterogeneity measured as  $I^2$  and the  
 302 alternatives ( $CV_{total}$ :  $r_{\text{spearman}} = 0.32$ , 95% CI = [0.24, 0.40],  $M_{total}$ :  $r_{\text{spearman}} = 0.33$ , 95% CI =  
 303 [0.25, 0.41]; Fig. 5). In cases of meta-analyses with  $I^2$  larger than 75% or smaller than 25%  
 304 (identified as large and small heterogeneity by conventional benchmarks<sup>10</sup>), the disagreement  
 305 between  $I^2$  and  $CV$ , as well as  $I^2$  and  $M$ , became even more pronounced (Fig. S5 – S7). In  
 306 contrast, a near-perfect agreement was observed between  $CV_{total}$  and  $M_{total}$ , as expected  
 307 ( $r_{\text{spearman}} = 1$ , 95% CI = [0.99, 1]; Fig. 5). Therefore, cross-meta-analysis (meta-scientific)  
 308 evidence suggests that the heterogeneity source measure  $I^2$  is not consistent with the  
 309 magnitude measures ( $CV_{total}$  and  $M_{total}$ ) for ecological and evolutionary data. We also found  
 310 that out of the 512 meta-analyses featuring medium to large  $I^2_{total}$  values (>50% based on  
 311 conventional guidelines), 80 had small  $CV_{total}$  (Fig. 5), indicating that more than 20% of the  
 312 large  $I^2_{total}$  values were caused by small sampling errors rather than larger amount of  
 313 heterogeneity. These findings emphasize the importance of considering multiple metrics to  
 314 obtain a holistic understanding of heterogeneity in meta-analyses (see **A pluralistic**  
 315 **framework**).

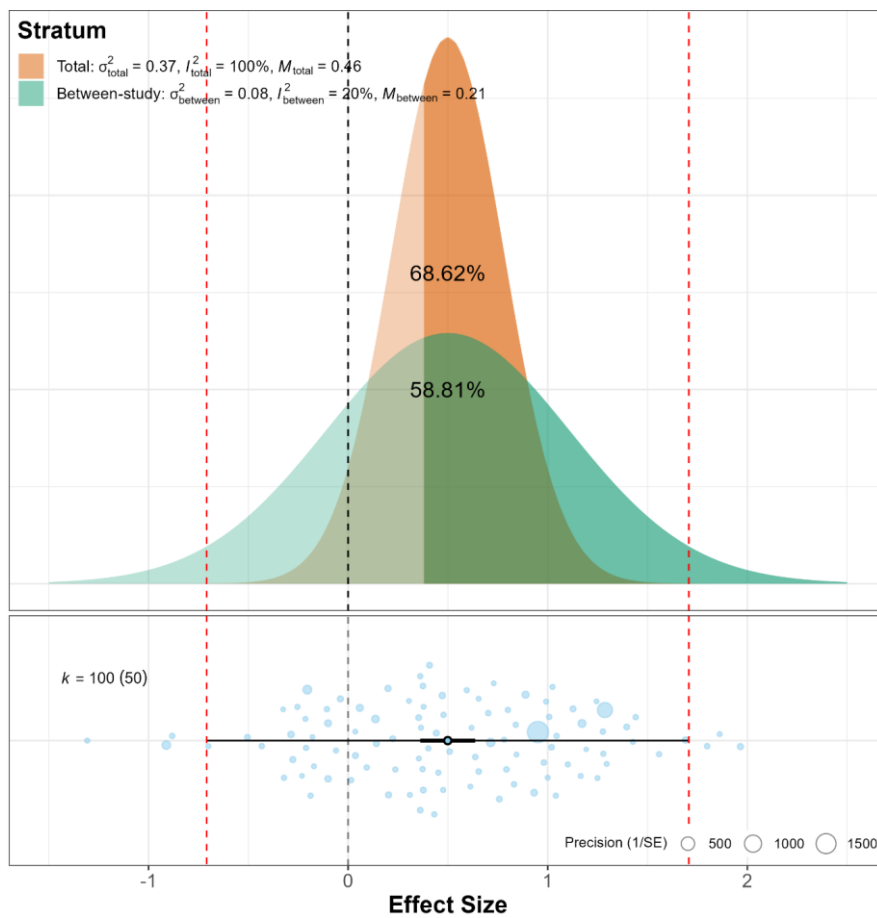


316  
 317 **Fig. 5:**  
 318 Disagreement (or agreement) between different heterogeneity metrics. For other details see Fig. 3.  
 319 The Spearman correlation estimates ( $r_{\text{spearman}}$ ) were: 0.32, 95% CI = [0.24, 0.40] for  $I^2_{total}$  and  $CV_{total}$ ,  
 320 0.33, 95% CI = [0.25, 0.41] for  $I^2_{total}$  and  $M_{total}$ , and 1, 95% CI = [0.99, 1] for  $M_{total}$  and  $CV_{total}$ .

321

322 **Prediction intervals and predictive distributions: visualising heterogeneity**

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



327

328 **Fig. 6:**

329 Example of how prediction intervals (PIs) combined with ‘prediction distributions’ (PDs) can be used  
330 to understand effect size heterogeneity and generality. Effect size data are simulated assuming two

331 effect sizes were collected from a total of  $n = 50$  studies, ( $k = 100$ ), with a  $\sigma_{between}^2 = 0.1$ ,  $\sigma_{within}^2 =$   
332 0.6 and an overall meta-analytic mean,  $\mu$ , of 0.5  
333 ([https://yefeng0920.github.io/heterogeneity\\_guide/](https://yefeng0920.github.io/heterogeneity_guide/)). Red dashed lines are the upper and lower  
334 95% PI, black dashed line the 'null' effect. The orchard plot <sup>30</sup> displays the overall meta-analytic  
335 mean, 95% confidence interval (CI) and 95% PI. The PDs were constructed using  $t$ -distribution with  $k$   
336 -1 degrees of freedom,  $\mu$  as location parameter, and total or between-study variance along with  
337 sampling variance of around  $\mu$  as scale parameter (see Equation 11). The percentage of effect sizes  
338 beyond a given threshold (i.e., the lower 95% CI) are provided.

339  
340 For example, consider a conservation intervention with a mean effect size (SMD) of -0.5 and  
341 95% PI of [-0.2 to -0.8]. This indicates that 95% of future interventions implemented in are  
342 predicted to decrease the conservation outcomes of interest by between 0.2 to 0.8 standard  
343 deviations. Unlike the point estimate of heterogeneity, such as  $\sigma_t^2$ , PIs offer an interval to  
344 inform the extent to which the focal effect can be generalized <sup>31</sup>. Under Equation 1, 95% PIs  
345 can be computed by <sup>7</sup>:

$$346 \quad 95\%PI = \mu \pm t_{0.975} \sqrt{\sigma_{between}^2 + \sigma_{within}^2 + SE[\mu]^2}, (11)$$

347 where  $t_{0.975}$  denotes the 97.5th percentile of a  $t$ -distribution (with  $k-1$  degrees of freedom <sup>32</sup>,  
348 where  $k$  is the number of sample size), and  $SE[\mu]$  denotes the standard error of the mean  
349 effect  $\mu$ .

350  
351 Despite their usefulness, PIs can create the illusion that all effect sizes within the upper and  
352 lower intervals are equally likely (Fig. 6; see also <sup>33</sup>). Therefore, statisticians have  
353 emphasised the importance of visualising the probability density to accurately capture the  
354 likelihood of each effect size within the intervals <sup>34,35</sup>. By considering the entire distribution  
355 of true effects while accounting for statistical noise, the predictive distribution (PD) offers a

356 more holistic measure of heterogeneity and generality. In the Bayesian framework, PDs,  
357 known as posterior distributions, are a natural part of the process, but even frequentist  
358 approaches can adopt PDs (sometimes referred to as “empirical Bayes”) to achieve similar  
359 aims. An advantage of the PD is its ability to calculate the probability that a true effect size  
360 exceeds a biologically or practically meaningful threshold although determining such a  
361 threshold usually requires domain-specific knowledge and expertise. The proportion of true  
362 effect sizes above a specific threshold could serve as a measure of evidence strength and  
363 generality<sup>16</sup>. Consider a case that 69% of effect sizes representing the efficacy of a  
364 conservation intervention are predicted to surpass a threshold value representing a practically  
365 significant effect (Fig.6, where we assumed the lower confidence limit representing the  
366 threshold). If assuming similar configurations of study contexts in the sampled future cases,  
367 we can infer that the intervention will achieve this benefit in 69% of future cases, with strong  
368 implications for policymaking.

369

### 370 **Modelling additional sources of heterogeneity**

371 In ecological and evolutionary datasets, complexity often arises from the inclusion of diverse  
372 species, temporal, and spatial variations<sup>3</sup>. Such complexity offers a unique opportunity for  
373 further disentangling heterogeneity. This can be achieved by embracing a flexible random-  
374 effects structure within the multilevel meta-analytic framework<sup>7,9</sup>. To illustrate this, we will  
375 show the principles of how to partition heterogeneity in datasets featuring multiple species  
376 (similar principles can be applied to those involving different temporal and spatial contexts).  
377 In the case of datasets encompassing multiple species, incorporating species-relevant  
378 random-effects terms into Equation 1 would lead to the phylogenetic multilevel meta-analytic  
379 model<sup>5,36</sup>:

380 
$$ES_{[i]} = \mu + u_{species[k]} + u_{phylogeny[k]} + u_{between[j]} + u_{within[i]} + e_{[i]}, (12)$$

381 where  $u_{s[k]}$  denotes the non-phylogenetic species random effect, which follows a normal  
 382 distribution with mean zero and variance  $\sigma_{species}^2$ ;  $u_{phylogeny[k]}$  denotes the phylogenetic  
 383 species random effect, which follows a multivariate normal distribution with mean zero and  
 384 variance-covariance matrix  $\sigma_{phylogeny}^2 \mathbf{A}$  (where  $\sigma_{phylogeny}^2$  is the phylogenetic species  
 385 variance, and  $\mathbf{A}$  is phylogenetic correlation matrix based on the distance between species on a  
 386 molecular-based phylogenetic tree).

387

388 With Equation 12 in hand, the total variance can be stratified at the phylogenetic and non-  
 389 phylogenetic species level ( $\sigma_{phylogeny}^2$  and  $\sigma_{species}^2$ ). Such stratification allows for the  
 390 assessment of the generality of a focal effect within these strata, as illustrated in the empirical  
 391 example below. Phylogenetic and non-phylogenetic species-level heterogeneity can be  
 392 measured using  $I_{phylogeny}^2$  and  $I_{species}^2$ , respectively <sup>5</sup>:

$$393 \quad I_{phylogeny}^2 = \frac{\sigma_{phylogeny}^2}{\sigma_{phylogeny}^2 + \sigma_{species}^2 + \sigma_{between}^2 + \sigma_{within}^2 + \bar{v}}, \quad (13)$$

$$394 \quad I_{species}^2 = \frac{\sigma_{species}^2}{\sigma_{phylogeny}^2 + \sigma_{species}^2 + \sigma_{between}^2 + \sigma_{within}^2 + \bar{v}}, \quad (14)$$

395 We derive the alternative stratified version of measures as follows:

$$396 \quad CV_{phylogeny} = \frac{\sigma_{phylogeny}}{|\mu|}, \quad (15)$$

$$397 \quad CV_{species} = \frac{\sigma_{species}}{|\mu|}, \quad (16)$$

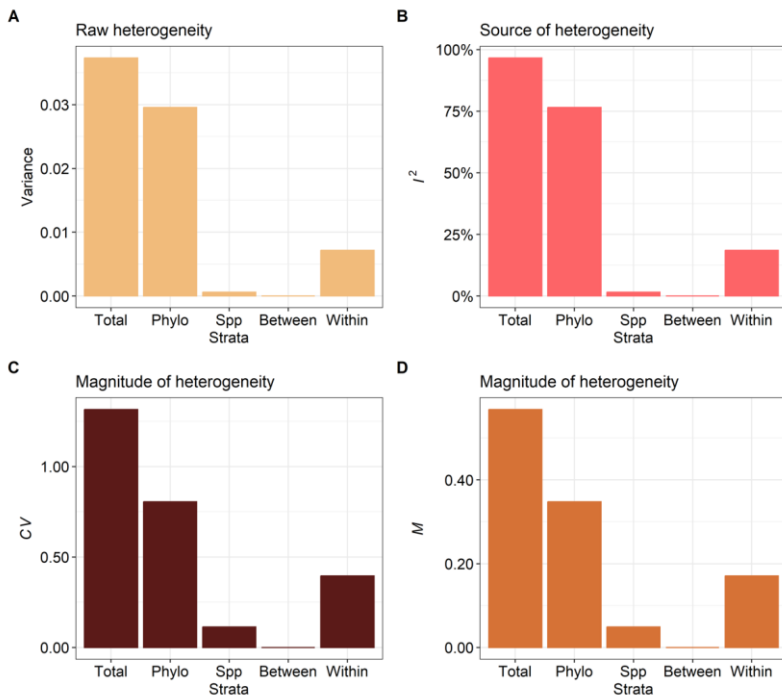
$$398 \quad M_{phylogeny} = \frac{\sigma_{phylogeny}}{\sigma_{phylogeny} + \sigma_{species} + \sigma_{between} + \sigma_{within} + |\mu|}, \quad (17)$$

$$399 \quad M_{species} = \frac{\sigma_{species}}{\sigma_{phylogeny} + \sigma_{species} + \sigma_{between} + \sigma_{within} + |\mu|}, \quad (18)$$

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

403

404 To illustrate the insights gained through these extended measures, we present an empirical  
405 example. We re-analysed a phylogenetic meta-analysis originally conducted by Risely et al.  
406 <sup>37</sup>. Our focus centres on a subset of this analysis, specifically examining the impact of  
407 infection status on the cost (e.g., movement capacity) of migratory animals. The data and  
408 code for replicating all calculations can be found at  
409 [https://yefeng0920.github.io/heterogeneity\\_guide/](https://yefeng0920.github.io/heterogeneity_guide/). Our re-analysis yielded three  
410 observations. Firstly,  $I_{total}^2 = 97\%$  exceeded the 75th percentile of the empirically derived  
411 heterogeneity distribution (Fig. 7 and Table S1). This suggests a high amount of  
412 heterogeneity according to the conventional benchmarks <sup>10</sup>. However, when we employed  
413 magnitude metrics to measure heterogeneity, they fell below between the 25th and 50th  
414 percentiles of the empirically derived heterogeneity distribution ( $CV_{total} = 1.3$  and  $M_{total} =$   
415  $0.6$ ). This discrepancy was attributed to the small typical sampling variance  $\bar{v}$ , which was  
416 found to be 0.001 in this case, underscoring  $I_{total}^2$ 's limitation of relying on  $\bar{v}$  to capture  
417 relative magnitude of heterogeneity. On the other hand, we emphasise that the proper  
418 interpretation of  $I_{total}^2$  is to use it to indicate the source of heterogeneity rather than the  
419 magnitude, as it represents the variance of the true effect in the context of the variance of the  
420 observed effect. For example,  $I_{total}^2 = 97\%$  suggests a heterogeneity can explain most (97%)  
421 of the variability in effect size (only 3% is explained by the sampling variance, or the  
422 heterogeneity is 32 times larger than that of statistical noise).



423

424 **Fig. 7:**

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

433 [https://yefeng0920.github.io/heterogeneity\\_guide/](https://yefeng0920.github.io/heterogeneity_guide/).

434

435 Secondly, the estimated mean effect was highly likely to be generalizable and replicable at  
436 the between-study- and species-context, if controlling for within-study experimental contexts  
437 (e.g., age, sex, outcomes). This is indicated by the stratification analysis that between-study  
438 level heterogeneity was extremely low, despite a large heterogeneity according to  
439 conventional benchmarks<sup>10</sup>. Traditional meta-analytic practices would overlook these  
440 valuable insights, potentially leading to erroneous conclusions. For example, random-effects  
441 meta-analysis shows that this dataset has high study-level heterogeneity ( $I^2_{total} = 96\%$ ; Fig. 5  
442 and Table S1). However, this amount of heterogeneity was not attributable to the study level  
443 but, rather, was mainly explained by the phylogenetic signal ( $I^2_{phylogeny} = 76\%$ ). The  
444 stratified version of PD also provided a clearer visual clue that the phylogenetic signal was  
445 the primary source of heterogeneity (Fig. 7).

446

## 447 **A pluralistic framework**

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

451 (1) Employing multilevel meta-analytic framework: Provided data allow, we strongly  
452 advocate for the use of a multilevel meta-analytic framework (Equation 1), as  
453 opposed to random-effects meta-analysis, for the modelling and stratification of  
454 heterogeneity. Additional random effects can be incorporated into Equation 1 as  
455 needed to further dissect heterogeneity. For example, the application of the  
456 phylogenetic multilevel meta-analytic model (Equation 12) allows for the  
457 disentanglement of species-specific heterogeneity.

458 (2) Quantification and stratification of pluralistic heterogeneity measures: We recommend  
459 transparently reporting all variance components, including typical sampling error



460 variances in the main text, supplementary tables, or figures (Figs. 6 and 7 and Table  
461 1). As such, pluralistic metrics can be computed using the formula above.  $I^2$ ,  $M$  (with  
462  $CV$  being derivable from  $M$ ), and their stratified versions should be reported as the  
463 default measures.  $PI$  or  $PD$  should also be reported to provide a visual identification  
464 of the heterogeneity information. These measures provide complementary  
465 information, for example, the source, magnitude, and visual clue of heterogeneity  
466 (examples see **Table 1**). We also provide parametric bootstrapping solutions to  
467 estimate the uncertainty (e.g., 95%CI) for each of the measures.

468 (3) Check the model parameter identifiability: When models incorporate many random  
469 effects, issues of parameter identifiability may arise, wherein unique variance  
470 estimates that maximize the likelihood function may not exist (see **Method**)<sup>39</sup>.  
471 Therefore, we recommend assessing whether variance components are all identifiable  
472 through means such as checking profile likelihood, before proceeding with  
473 heterogeneity quantification and stratification.

474 (4) Carefully interpret heterogeneity measures: It is crucial to interpret both total and  
475 stratified heterogeneity to evaluate variation in effect sizes, aiding in the examination  
476 of general rules in the fields of ecology and evolution (see a case study in **Modelling  
477 additional sources of heterogeneity**). However, neither the conventional benchmarks  
478 (25, 50, and 75% as small, moderate and high heterogeneity<sup>10</sup>) nor those of  
479 empirically derived distributions (Table 1 and Fig. 3) are currently suitable for  
480 informing interpretation. Nevertheless, the empirically derived distribution can be  
481 employed to interpret heterogeneity within the context of existing ecological and  
482 evolutionary meta-analyses.

483

484 We argue that ecologists and evolutionary biologists should treat heterogeneity and the meta-  
485 analytic mean effect size with equal importance. We provide a user-friendly tutorial equipped  
486 with a set of R functions to streamline the qualification, stratification, and interpretation of  
487 heterogeneity [https://yefeng0920.github.io/heterogeneity\\_guide/](https://yefeng0920.github.io/heterogeneity_guide/), empowering ecologists and  
488 evolutionary biologists to discern generality.  
489

490 Table 1

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

Types	Metrics	Interpretation and examples	Empirically derived benchmark <sup>1</sup>
Test statistic	$Q$	Null-hypothesis test. Statistical test of heterogeneity in effect sizes.	Not applicable
Unstandardisation	$\sigma^2$	Absolute magnitude measure of heterogeneity. Variance (square of standard deviation) of the meta-analytic mean effect ( $\sigma_{total}^2$ ) and its stratification at between- and within-study contexts ( $\sigma_{between}^2$ and $\sigma_{within}^2$ ).	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 <i>Zr</i> ; 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, $\sigma_{total}^2 = 95\%$ denotes that 95% of variation is the result of nuisance heterogeneity (i.e., differences in contexts). $\sigma_{between}^2 = 80\%$ and $\sigma_{within}^2 = 15\%$ 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 <i>Zr</i> ; 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 especially so for CV and M

Should discuss with Shinichi

**Commented [YY2R1]:** Good point

Mean-standardization	<i>CV</i>	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, 1.3 for uncommon measures.
Variance-mean-standardization	<i>M</i>	Heterogeneity magnitude measure. Variance expressed as the proportion of the mean effect and a transformation of <i>CV</i> 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

495 <sup>1</sup>The distributions and percentiles could be underestimated if publication bias existed.

## 497 **Methods**

### 498 **Meta-analysis database**

499 The ecological and evolutionary database used in this study were originally compiled by  
500 Costello<sup>18</sup>, O'Dea<sup>17</sup>, and their colleges. They conducted a systematic search for meta-  
501 analysis papers published in ecological journals, including those from the Ecological Society  
502 of America and journals of the British Ecological Society. Additionally, they supplemented  
503 the database with high-profile journals, such as Nature, and Science. Their systematic search  
504 yielded 522 meta-analysis datasets. We dropped meta-analysis datasets that could not achieve  
505 convergence when fitted to the multilevel model. Convergence could not be reached for ten  
506 meta-analysis datasets, even after adjusting key parameters of the iterative methods to  
507 maximize the log likelihood function (see below for details). Therefore, our database  
508 contained 512 meta-analysis datasets encompassing 17,770 primary studies and 109,495  
509 effect size estimates. On average, each meta-analysis dataset included 240 effect size  
510 estimates sourced from 40 studies, with median values of 64 and 23, respectively.

511

### 512 **Stratification of hierarchical meta-analytic data**

513 In this section, we elucidate the theoretical background behind employing a three-level meta-  
514 analytic approach to stratify datasets characterized by three-level hierarchical structure as  
515 outlined above. Note that the stratification of heterogeneity can be further extended to data  
516 structures with more than four strata as necessary (see a case study in **Model additional**  
517 **source heterogeneity**). In the first-stage modelling procedure, the true (population) effect  
518 size  $\mu_{between[j]}$  of  $j$ -th study is modelled using a normal distribution with expectation  $\mu$  and  
519 variance  $\sigma_{between}^2$ , where  $\mu$  is the population mean effect or overall effect and  $\sigma_{between}^2$   
520 denotes the extent to which  $\mu_{between[j]}$  deviates from the overall effect  $\mu$ <sup>24,40</sup>. Moving to the  
521 second-stage modelling procedure, the  $i$ -th effect size  $\mu_{within[i]}$  within  $j$ -th study is modelling

522 using a normal distribution with expectation  $\mu_{between[j]}$  and variance  $\sigma_{within}^2$ , where  $\sigma_{within}^2$   
 523 represents the extent to which within-study effect  $\mu_{within[i]}$  deviates from between-study  
 524 effect  $\mu_{between[j]}$ <sup>24,40</sup>. In the third-stage modelling procedure, the effect size estimate  $ES_{[i]}$  of  
 525  $\mu_{within[i]}$  is modelled using a normal distribution with expectation  $\mu_{within[i]}$  and sampling  
 526 error variance  $v_{[i]}$ . This multilevel modelling framework provides a general way to  
 527 decompose the variance of effect sizes into different strata, for example between- and within-  
 528 study levels.

529  
 530 From the implementation perspective, effect size estimate  $ES_{[i]}$  is not sequentially modelled  
 531 through the three-stage process but rather directly modelled from the overarching distribution  
 532 with an expectation  $\mu$  and variance-covariance matrix  $VCV$ <sup>24,40</sup>:

$$533 \begin{bmatrix} \sigma_{between}^2 + \sigma_{within}^2 + v_{[1]} & \cdots & \sigma_{between}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{between}^2 & \cdots & \sigma_{between}^2 + \sigma_{within}^2 + v_{[k]} \end{bmatrix}, (19)$$

534 The meta-analytic model specified with the variance-covariance matrix  $VCV$  is referred to as  
 535 the multilevel meta-analytic model (Equation 1).  $VCV$  can be reparametrized as a compound  
 536 symmetry random-effects structure within the framework of multivariate meta-analytic model  
 537 <sup>40,41</sup>.

$$538 \begin{bmatrix} \sigma_{total}^2 + v_{[1]} & \cdots & \rho\sigma_{total}^2 \\ \vdots & \ddots & \vdots \\ \rho\sigma_{total}^2 & \cdots & \sigma_{total}^2 + v_{[k]} \end{bmatrix}, (20)$$

539 where  $\sigma_{total}^2 = \sigma_{between}^2 + \sigma_{within}^2$  is the total variance in effect sizes and  $\rho =$   
 540  $\sigma_{between}^2 / \sigma_{total}^2$  denotes intraclass correlation coefficient. We used the *rma.mv()* function  
 541 from the *metafor* package<sup>42</sup> to fit all 512 meta-analysis datasets to the three-level meta-  
 542 analytic model (Equation 1). We employed restricted maximum likelihood (REML) as the  
 543 variance estimator and the quasi-Newton method as the optimizer to maximize the likelihood

544 function over variance estimation ( $\sigma_{between}^2$  and  $\sigma_{within}^2$ ), with a threshold of  $10^{-8}$ , a step  
545 length of 1, and a maximum iteration limit of 1000. All models successfully converged under  
546 these settings. We confirmed the identifiability of variance estimation ( $\sigma_{between}^2$  and  $\sigma_{within}^2$ )  
547 by checking their likelihood profiles. The R code for model fitting can be accessed at the  
548 [https://github.com/Yefeng0920/heterogeneity\\_ecoevo](https://github.com/Yefeng0920/heterogeneity_ecoevo).

549

### 550 **Extended heterogeneity metrics**

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

$$557 \quad CV_{total}^2 = \frac{\sigma_{total}^2}{\mu^2}, \quad (21)$$

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

$$560 \quad CV_{between}^2 = \frac{\sigma_{between}^2}{\mu^2}, \quad (22)$$

$$561 \quad CV_{within}^2 = \frac{\sigma_{within}^2}{\mu^2}, \quad (23)$$

562 Following the same principle,  $M_{total}^2$  can be obtained <sup>11</sup>:

$$563 \quad M_{total}^2 = \frac{\sigma_{total}^2}{\sigma_{total}^2 + \mu^2}, \quad (24)$$

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

566 
$$M_{between}^2 = \frac{\sigma_{between}^2}{\sigma_{total}^2 + \mu^2}, (25)$$

567 
$$M_{within}^2 = \frac{\sigma_{within}^2}{\sigma_{total}^2 + \mu^2}, (26)$$

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



572 **Data availability**

573 The data needed to reproduce the analyses and figures are archived GitHub repository  
574 [https://github.com/Yefeng0920/heterogeneity\\_ecoevo/tree/main](https://github.com/Yefeng0920/heterogeneity_ecoevo/tree/main), and will be deposited at  
575 Zenodo after acceptance.

576 **Code availability**

577 The scripts needed to reproduce the analyses and figures are archived GitHub repository  
578 [https://github.com/Yefeng0920/heterogeneity\\_ecoevo/tree/main](https://github.com/Yefeng0920/heterogeneity_ecoevo/tree/main), and will be deposited at  
579 Zenodo after acceptance.

580

581 **References**

- 582 1 Lawton, J. H. Are there general laws in ecology? *Oikos*, 177-192 (1999).
- 583 2 Spake, R. *et al.* Improving quantitative synthesis to achieve generality in ecology. *Nature*  
584 *Ecology & Evolution*, 1-11 (2022).
- 585 3 Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of  
586 research synthesis. *Nature* **555**, 175-182 (2018).
- 587 4 Martin, P. A. *et al.* Flexible synthesis can deliver more tailored and timely evidence for  
588 research and policy. *Proceedings of the National Academy of Sciences* **120**, e2221911120  
589 (2023).
- 590 5 Nakagawa, S. & Santos, E. S. Methodological issues and advances in biological meta-analysis.  
591 *Evolutionary Ecology* **26**, 1253-1274 (2012).
- 592 6 Noble, D. W. *et al.* Meta-analytic approaches and effect sizes to account for 'nuisance  
593 heterogeneity' in comparative physiology. *Journal of Experimental Biology* **225**, jeb243225  
594 (2022).
- 595 7 Yang, Y., Macleod, M., Pan, J., Lagisz, M. & Nakagawa, S. Advanced methods and  
596 implementations for the meta-analyses of animal models: Current practices and future  
597 recommendations. *Neuroscience & Biobehavioral Reviews*, 105016 (2022).
- 598 8 Senior, A. M. *et al.* Heterogeneity in ecological and evolutionary meta - analyses: its  
599 magnitude and implications. *Ecology* **97**, 3293-3299 (2016).
- 600 9 Nakagawa, S., Yang, Y., Macartney, E. L., Spake, R. & Lagisz, M. Quantitative evidence  
601 synthesis: a practical guide on meta-analysis, meta-regression, and publication bias tests for  
602 environmental sciences. *Environmental Evidence* **12**, 8, doi:10.1186/s13750-023-00301-6  
603 (2023).
- 604 10 Higgins, J. P., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-  
605 analyses. *BMJ* **327**, 557-560 (2003).
- 606 11 Cairns, M. & Prendergast, L. A. On ratio measures of heterogeneity for meta - analyses.  
607 *Research Synthesis Methods* **13**, 28-47 (2022).
- 608 12 Rucker, G., Schwarzer, G., Carpenter, J. R. & Schumacher, M. Undue reliance on I2 in assessing  
609 heterogeneity may mislead. *BMC medical research methodology* **8**, 1-9 (2008).
- 610 13 Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta - analysis. *Statistics in*  
611 *medicine* **21**, 1539-1558 (2002).
- 612 14 IntHout, J., Ioannidis, J. P., Rovers, M. M. & Goeman, J. J. Plea for routinely presenting  
613 prediction intervals in meta-analysis. *BMJ open* **6**, e010247 (2016).
- 614 15 Borenstein, M., Higgins, J. P., Hedges, L. V. & Rothstein, H. R. Basics of meta - analysis: I2 is  
615 not an absolute measure of heterogeneity. *Research synthesis methods* **8**, 5-18 (2017).
- 616 16 Mathur, M. B. & VanderWeele, T. J. New metrics for meta - analyses of heterogeneous  
617 effects. *Statistics in Medicine* **38**, 1336-1342 (2019).
- 618 17 O'Dea, R. E. *et al.* Preferred reporting items for systematic reviews and meta - analyses in  
619 ecology and evolutionary biology: a PRISMA extension. *Biological Reviews* **96**, 1695-1722  
620 (2021).
- 621 18 Costello, L. & Fox, J. W. Decline effects are rare in ecology. *Ecology*, e3680 (2022).
- 622 19 Noble, D. W., Lagisz, M., O'dea, R. E. & Nakagawa, S. Nonindependence and sensitivity  
623 analyses in ecological and evolutionary meta - analyses. *Molecular Ecology* **26**, 2410-2425  
624 (2017).
- 625 20 Yang, Y. *et al.* Robust point and variance estimation for ecological and evolutionary meta-  
626 analyses with selective reporting and dependent effect sizes. *EcoEvoRxiv*,  
627 doi:<https://doi.org/10.32942/X20G6Q> (2023).

628 21 Viechtbauer, W. & López-López, J. A. Location-scale models for meta-analysis. *Research*  
629 *synthesis methods* **13**, 697-715 (2022).  
630 22 Cochran, W. G. The combination of estimates from different experiments. *Biometrics* **10**, 101-  
631 129 (1954).  
632 23 Takkouche, B., Cadarso-Suarez, C. & Spiegelman, D. Evaluation of old and new tests of  
633 heterogeneity in epidemiologic meta-analysis. *American journal of epidemiology* **150**, 206-  
634 215 (1999).  
635 24 Cheung, M. W.-L. Modeling dependent effect sizes with three-level meta-analyses: a  
636 structural equation modeling approach. *Psychological Methods* **19**, 211 (2014).  
637 25 Hansen, T. F., Pélabon, C. & Houle, D. Heritability is not evolvability. *Evolutionary Biology* **38**,  
638 258-277 (2011).  
639 26 Nakagawa, S. *et al.* Meta-analysis of variation: ecological and evolutionary applications and  
640 beyond. *Methods in Ecology and Evolution* **6**, 143-152 (2015).  
641 27 Dochtermann, N. A. & Royauté, R. The mean matters: going beyond repeatability to interpret  
642 behavioural variation. *Animal Behaviour* **153**, 147-150 (2019).  
643 28 Yang, Y. *et al.* Publication bias impacts on effect size, statistical power, and magnitude (Type  
644 M) and sign (Type S) errors in ecology and evolutionary biology. *BMC biology* **21**, 1-20 (2023).  
645 29 Richter, S. H. Systematic heterogenization for better reproducibility in animal  
646 experimentation. *Lab animal* **46**, 343-349 (2017).  
647 30 Nakagawa, S. *et al.* The orchard plot: Cultivating a forest plot for use in ecology, evolution,  
648 and beyond. *Research Synthesis Methods* **12**, 4-12 (2021).  
649 31 van Aert, R. C., Schmid, C. H., Svensson, D. & Jackson, D. Study specific prediction intervals  
650 for random-effects meta-analysis: A tutorial: Prediction intervals in meta-analysis.  
651 *Research synthesis methods* **12**, 429-447 (2021).  
652 32 Knapp, G. & Hartung, J. Improved tests for a random effects meta-regression with a single  
653 covariate. *Statistics in medicine* **22**, 2693-2710 (2003).  
654 33 Bishop, J. & Nakagawa, S. Quantifying crop pollinator dependence and its heterogeneity  
655 using multi-level meta-analysis. *Journal of Applied Ecology* **58**, 1030-1042 (2021).  
656 34 Jackson, C. H. Displaying uncertainty with shading. *The American Statistician* **62**, 340-347  
657 (2008).  
658 35 Barrowman, N. J. & Myers, R. A. Raindrop plots: a new way to display collections of  
659 likelihoods and distributions. *The American Statistician* **57**, 268-274 (2003).  
660 36 Cinar, O., Nakagawa, S. & Viechtbauer, W. Phylogenetic multilevel meta-analysis: A  
661 simulation study on the importance of modelling the phylogeny. *Methods in Ecology and*  
662 *Evolution* **13**, 383-395 (2022).  
663 37 Risely, A., Klaassen, M. & Hoyer, B. J. Migratory animals feel the cost of getting sick: A meta-  
664 analysis across species. *Journal of Animal Ecology* **87**, 301-314 (2018).  
665 38 Voelkl, B. *et al.* Reproducibility of animal research in light of biological variation. *Nature*  
666 *Reviews Neuroscience*, 1-10 (2020).  
667 39 Raue, A. *et al.* Structural and practical identifiability analysis of partially observed dynamical  
668 models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923-1929 (2009).  
669 40 Van den Noortgate, W., López-López, J. A., Marín-Martínez, F. & Sánchez-Meca, J. Three-level  
670 meta-analysis of dependent effect sizes. *Behavior research methods* **45**, 576-594 (2013).  
671 41 Cheung, M. W.-L. A guide to conducting a meta-analysis with non-independent effect sizes.  
672 *Neuropsychology review* **29**, 387-396 (2019).  
673 42 Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *Journal of*  
674 *statistical software* **36**, 1-48 (2010).

675

676 **Acknowledgements**

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

681

682 **Author contributions**

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

690

691 **Competing interests**

692 All authors declare no competing interests.

693

694 **Additional information**

695 Supplementary materials will be available at the online version.