Comparing ecological and evolutionary variability within datasets

Raphaël Royauté, Ned A Dochtermann

HAL Id: hal-03955506
https://hal.inrae.fr/hal-03955506

Submitted on 25 Jan 2023

Comparing ecological and evolutionary variability within datasets

1   Raphaël Royauté [a,b*] and Ned A. Dochtermann [a]

2   [a] Department of Biological Sciences; North Dakota State University, Fargo, ND, USA

3   [b] Current address: Movement Ecology Group, Senckenberg Biodiversity and Climate

4   Research Centre (SBiK-F), Frankfurt, Germany

5   [*] corresponding author: raphael.royaute@gmail.com

6   **ORCID IDs:** 0000-0002-5837-633X; 0000-0002-8370-4614

7

8

9   Running Head: Comparing variation within datasets

10

13

14

ABSTRACT

Many key questions in evolutionary ecology require the use of variance ratios such as heritability, repeatability, and individual resource specialization. These ratios allow researchers to understand how phenotypic variation is structured into genetic and non-genetic components, to identify how much organisms vary in the resources they use or how functional traits structure species communities. Understanding how evolutionary and ecological processes differ among populations and environments therefore often requires the comparison of these ratios across groups (i.e. populations, sexes, species). Inference based on comparisons of ratios can be limited, however. Variance ratios can remain the same across group despite very different values in the numerator and denominator variances. Moreover, evolutionary ecologists are most often interested in differences in specific variance components among groups rather than in differences in variance ratios *per se*. Recommendations for how to infer whether groups differ in variance are not clear in the literature. Using simulations, we show how questions regarding the estimation of variance components and their differences among groups can be answered with Linear Mixed Models (LMMs). Frequentist and Bayesian frameworks have similar abilities to identify differences in variance components. However, variance differences at higher levels of organization can be difficult to detect with low sample sizes. We provide tools to conduct power analyses to determine the appropriate sample sizes necessary to detect differences in variance of a given magnitude. We conclude by supplying guidelines for how to report and draw inferences based on the comparisons of variance components and variance ratios

SIGNIFICANCE STATEMENT

37   Many critical questions in ecology and evolution use variance ratios, such as repeatability,

38   heritability, or individual resource specialization, to make inferences about ecological and

39   evolutionary processes. In many cases these inferences rely on the comparison of variance

40   ratios among datasets (populations, sexes, or environments). In this article, we show that

41   current approaches of drawing inferences about group differences from comparisons of

42   ratios are inappropriate because ratios can differ due to differences in the numerator,

43   denominator, or both. We investigated how questions regarding differences in variance

44   ratios and constituent variance components can be evaluated using Linear Mixed Model

45   approaches (LMMs) and provide guidance for appropriate sampling schemes under

46   different scenarios and discuss common pitfalls associated with estimation of differences in

47   variance component among datasets.

48

49   Running Head: Comparing variation within datasets

50

51   Keywords: Repeatability, animal personality, individual variation, mixed models, individual

52   niche specialization, functional traits

53

Declarations

Funding

This study was funded by NSF IOS-1557951 (to NAD) and the Department of Biological Sciences at North Dakota State University.

Conflicts of interest/Competing interests

The authors declare no conflict of interest

Availability of data and material

All code and data for simulations is available on the Open Science Framework's project for this article: https://osf.io/5aw42/

Code availability

All code and data for simulations is available on the Open Science Framework's project for this article: https://osf.io/5aw42/

Author contribution

Each author contributed equally to the design, analysis and writing of the manuscript.

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Not applicable

INTRODUCTION

Our understanding of many evolutionary and ecological processes is underpinned by an estimation of variance ratios (Table 1). For example, the reporting of repeatability has become pervasive in behavioral studies as it summarizes the amount of variation in behavior attributable to differences among individuals. Informally these differences among individuals can be thought of as differences in their average behaviors. Repeatability then can be interpreted as how much of the overall variation is attributable to individual differences

Use of variance ratios like repeatability spans a broad swath of evolutionary ecology (Table 1). This includes the most well-known variance standardized ratio: heritability, and extends to interest in community ecology regarding the distribution of functional trait variation expressed within versus among populations or species (Violle et al. 2012).

While useful for understanding the relative magnitude of variation, variance ratios can be highly misleading when compared between groups (Houle 1992; Wilson 2018). Comparisons of variance ratios are only narrowly interpretable because these ratios can differ when numerators differ, when denominators differ, or when both differ. Indeed, variance ratios can be equal despite having different numerators and denominators values. Put another way, differences between groups in ratios like repeatability are not informative as to absolute differences in the magnitudes of variation observed.

94 **Table 1** Examples variance ratios found in the the ecological and evolutionary literature

Discipline	Variance ratio	Definition	Description	References
Quantitative Genetics	*Heritability*	$h^2 = Va / Vp$	The proportion of variation attributable to additive genetic variance (Va)	(Mousseau and Roff 1987)
Behavioral Ecology	*Adjusted Repeatability*	$R_A = Vi / (Vi + Vw)$	The proportion of variation attributable to among-individual differences (Vi) relative to either the total variation (Vi+Vf+Vw) or after adjusting for fixed-effects (Vi+Vw)	(Lessells and Boag 1987)
	Unadjusted Repeatability	$R_U = Vi / (Vi + V_f + Vw)$		(Nakagawa and Schielzeth 2010)
Ecology	*Individual Niche Specialization*	$S = WIC / TNW$	The proportion of variation attributable to within-individual preference in niche (WIC) (usually expressed as standard deviations)	(Bolnick et al. 2002)
Community Ecology	*T-ratios*	$T_{IP/IC} = V_{IP} / V_{IC}$	The proportion of variation attributable to within-population variance (V_{IP}) relative to the community variance (V_{IC})	(Violle et al. 2012)
		$T_{IC/IR} = V_{IC} / V_{IR}$	The proportion of variation attributable to community variance (V_{IC}) relative to the regional pool variance (V_{IR})	(Violle et al. 2012)

95

96 Legend: *Va*: additive genetic variance in a trait, Vi: among-individual variance in trait, *Vw*: within-individual (i.e. residual)
97 variance in a trait, WIC: within-individual variance in niche preference, *TNW*: Total niche width, T_{IP}: total amount of trait
98 variation in a community, V_{IP}: within-population variance in trait, V_{IC}: community variance in trait, V_{IR}: regional pool variance

99   To further illustrate the inferential limits of variance ratios, consider the following

100   scenario: researchers are studying the behaviors and dietary habits of two populations of

101   the mythical Dahu (*Dahu desterus*; Fig. 1A) at different elevations. These elusive creatures

102   have shorter hind-legs on their left side, thus only allowing for clockwise movement

103   (Chartois and Claudel 1945; Jacquat 1995). While measuring aggressive interactions,

104   researchers find no differences in means between populations and similar behavioral

105   repeatabilities ($\tau = 0.8$; Fig. 1B). Put another way, the same relative amount of variation is

106   attributable to individuals in each population. The researchers notice, however, that there

107   are large differences in the among- and within-individual variances of each population. Had

108   researchers only examined repeatabilities and mean differences they would

109   inappropriately conclude that the populations are behaviorally equivalent. Instead, the

110   actual variance estimates reveal that individuals from the high-altitude population are very

111   distinct from one another in their aggressive tendencies while, at low-altitude, individuals

112   show little departure from the population average (Fig. 1B, C).

113   These researchers are also curious as to whether the harsher climate at the top of

114   the mountain range leads to a narrower dietary breadth. Researchers predict that

115   individual resource specialization will be higher in the low elevation population, as *D.*

116   *desterus* have more food options to choose from. To the researcher's surprise, they find

117   much higher individual resource specialization in the high-altitude population: $S_1 = 0.2$, $S_2 =$

118   0.8. Upon examining the specific values of among- and within-individual variation in niche,

119   they find that these differences are a result of the high elevation population having a much

120   narrower total niche width (Fig. 1D) while the within-individual variation in niche

121   preference is equal between populations.

122

123
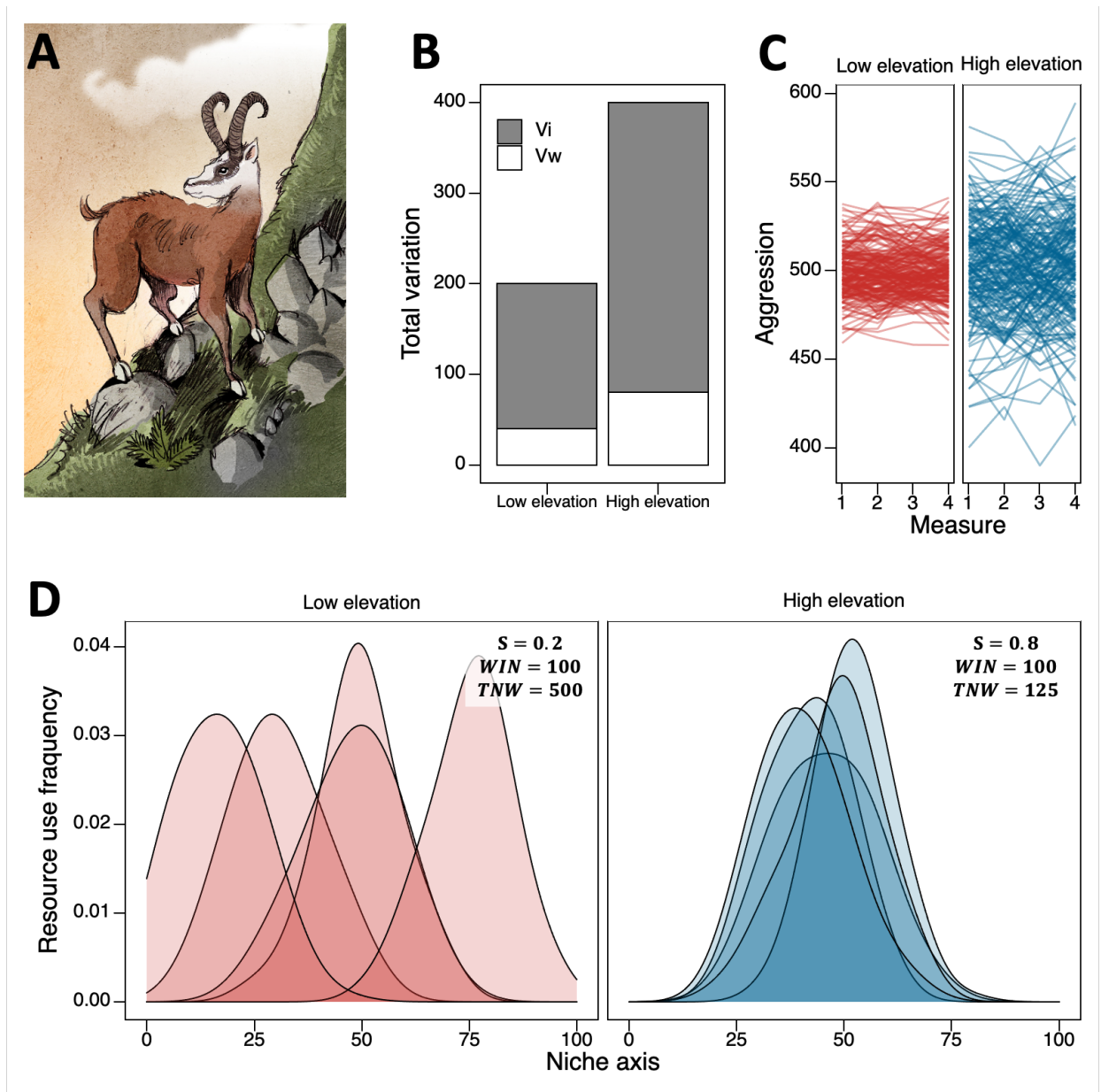
Fig. 1 Reliance on variance ratios can lead to misleading inferences. (A) The elusive Dahu (*Dahu dexterus*) in its natural environment. (B) Two populations of Dahus living at different elevations do not differ in their repeatability of aggressive interactions (τ). (C) By plotting the individual aggression scores over the course of multiple measurements, it is clear that individuals are more distinct in their aggressive behavioral strategies at high elevation. This inference cannot be made by investigating repeatability alone. (D) The two population have very different resource specialization indices (S). A more accurate inference is that individuals do not differ in niche width (*WIN*), it is instead the total niche wdith (*TNW*) that is narrower in the high-alttitude population. Code available here: https://osf.io/5aw42/.
Illustration: Philippe Semeria (CC BY 3.0 license)

134   This means that it is the difference in diet preference among individuals that drives

135   the difference between the two populations. With more varied resources available at low

136   elevation, each individual can specialize along the total niche axis, yet the breadth of diet

137   preference within-individuals is the same between populations.

138   For both traits, exclusive reliance on ratios would have led to either inappropriate

139   or incomplete inferences (i.e. inappropriately concluding behavioral equivalence and

140   incompletely recognizing the basis of differences in apparent specialization). Due to these

141   problems with interpretations of variance ratios (Houle 1992; Dochtermann and Royauté

142   2019), what would be of greater use to researchers is to instead evaluate differences in

143   specific variance components.

144   *A statistical framework for comparing variance components*

145   The statistical procedures necessary for the estimation of variance components and ratios

146   within a single population have been the subject of much attention (e.g. mixed models for

147   repeatability: Dingemanse and Dochtermann 2013; animal models for heritability: Wilson

148   et al. 2010; individual niche specialization:Bolnick et al. 2002; Coblentz et al. 2017;

149   functional trait variation: Nakagawa and Schielzeth 2012; Violle et al. 2012; Carmona et al.

150   2016). There is also a long history in quantitative genetics regarding the comparison of

151   variances and *co*variance structures among groups (Shaw 1991; Arnold and Phillips 1999;

152   Roff 2002; Roff et al. 2012; Aguirre et al. 2014). Unfortunately, these quantitative genetic

153   approaches have been poorly disseminated across fields (but see (Dochtermann and Roff

154   2010; White et al. 2020). Here we describe and investigate methods for detecting

155   differences in variance components amongst groups. Specifically, we compare the strength

156   and weaknesses of three statistical approaches: comparison of confidence intervals, model

157   comparison with AIC, and Bayesian estimation of the difference in variance components.

158   While this selection of methods encompasses very different philosophical approaches to

159   data analysis, all three are routinely used in the estimation of repeatability and other ratios.

160   We consider a scenario where a phenotypic attribute, y, is measured repeatedly for

161   individual organisms occupying one of two different environments (E1 and E2) and in

162   which variation occurs among and within-individuals (V_I and V_W respectively). In the

163   following sections we focus on differences in individual variation and repeatability. Note,

164   however, that this scenario can also be expanded to the comparison of diet specialization

165   for individuals occupying different environments or how functional traits vary among and

166   within species in two different environments.

167   An easy way to compare these variance components and their ratios ($\tau = V_I/(V_I +$

168   $V_W)$) is to estimate the variance components for each environment in separate statistical

169   models. We can then test for differences in variances and ratios by environment based on

170   whether estimate confidence intervals overlap or not. While straightforward, this method

171   suffers from two key limitations. First, basing inference on the overlap of 95 % confidence

172   intervals is overly conservative (Barr 1969), especially when sample size is low. It is

173   instead whether the confidence interval for the *difference* in variances excludes 0 that is

174   relevant for drawing inferences. This difference cannot be directly estimated from the

175   approach we have described. However, statistical significance can still be assessed by

176   comparing the overlap of the 83% confidence intervals for variance components, a

177   threshold that provides a better approximation for an $\alpha = 0.05$ for the null hypothesis of no

178   difference (Schenker and Gentleman 2001; Austin and Hux 2002; MacGregor-Fors and

179   Payton 2013; Hector 2021). Second, by estimating variance components in separate

180   statistical models, the hierarchical structure of the data, i.e. the variance components

181   nested within the environments, has been broken. As a result, potential average differences

182   in the traits of interest are not appropriately tested.

183   Instead, we suggest that a more appropriate procedure would be the use of a Linear

184   Mixed Model (LMM) where the among- and within-individual variance is estimated for

185   each environment within the same statistical model. This statistical model can be described

186   by the following equation:

187    $y_{ij} = \beta_0 + \beta_1 Environment + ID_{0i} + e_{0ij}$ (equation 1)

188    $ID_{0i} \sim MVN(0, \Omega_{ID}); \quad \Omega_{ID} = \begin{bmatrix} V_{ID}\,E_1 & 0 \\ 0 & V_{ID}\,E_2 \end{bmatrix}$

189    $e_{0ij} \sim MVN(0, \Omega_e); \quad \Omega_e = \begin{bmatrix} V_e\,E_1 & 0 \\ 0 & V_e\,E_2 \end{bmatrix}$

190   where y_{ij} describes the phenotypic traits for the ith individual and jth observation. ID_{0i}, is

191   the deviation from an overall intercept, β_0, for the ith individual. β_1 represents the

192   regression coefficient for the fixed effect of environment (here a contrast coefficient). The

193   random intercepts and residual variance (e_{0ij}) both follow a multivariate normal

194   distribution, and Ω_{ID} and Ω_e, are the variance-covariance matrices at the among- and

195   within-individual levels respectively.

196   The diagonal elements of these matrices represent the among- and within-

197   individual variances in each environment: E_1 and E_2. The off-diagonal elements represent

198   the cross-environment correlation (set to 0 if individuals are only ever evaluated in one of

199   the two environments). This formulation has the advantage of allowing considerable

200   flexibility in the specification of the statistical models considered (Dingemanse and

201 Dochtermann 2013). LMMs are now available for most statistical software and their

202 generalized extensions can accommodate non-normal error distributions (Table 2).

203   Upon fitting LMMs, several methods are then available to determine whether a

204 variance ratio or components of the ratio differ by environment. Specific hypotheses of

205 which variance component differs across environment can be easily tested via model

206 comparison. For example, a model where only the among-individual variance differs by

207 environment can be compared to a null model where the among and within- individual

208 variance are kept constant across developmental environments (Royauté et al. 2019).

209 These models can be estimated within a frequentist framework via restricted maximum

210 likelihood or a Bayesian framework and suitable decision criteria can be used to determine

211 which model best fits the data. In the case of restricted maximum likelihood estimation, it is

212 also possible to use likelihood ratio tests to compare these models. Note however that the

213 proper degrees of freedom to apply to each model is unclear and additional care should be

214 taken when using this method (Pinheiro and Bates 2006; Santostefano et al. 2016). We

215 recommend calculating these degrees of freedom by considering each variance component

216 as a full parameter for more conservative testing (see also the tutorial in ESM3).

217 **Table 2** Packages and softwares allowing to test for differences in variance components using Linear Mixed Models (LMM) along with
218 parameter estimation method (maximum likelihood (ML), restricted maximum likelihood (REML), hierachical likelihood (H-ML) or
219 Bayesian framework) and inference method (Likelihood Ratio tests (LRT), AIC, bootstrapping or credible interval for ΔV). This list is not
220 comprehensive and is instead based on widely-used commercial softwares and R packages
221

Package or software	Free or commercial	Estimation	Testing method	Among-unit variance by group	Residual variance by group	Distributions handled	Comments	Reference
ASREmL	Commercial	ML/REML	LRT, AIC, bootstrapping	Yes	Yes	Gaussian		(Gilmour et al. 2015)
SAS	Commercial	ML/REML	LRT, AIC, bootstrapping	Yes	Yes	Gaussian, Poisson, Binomial …		SAS Institute Inc.
nlme	Free	ML/REML	LRT, AIC, bootstrapping	Yes	Yes	Gaussian		(Pinheiro and Bates 2006)
lme4	Free	ML/REML	LRT, AIC, bootstrapping	Yes	No	Gaussian, Poisson, Binomial …		(Bates et al. 2015)
glmmTMB	Free	ML/REML	LRT, AIC, bootstrapping	Yes	Yes	Gaussian, Poisson, Binomial …		(Brooks et al. 2017)
hglm	Free	H-ML	LRT, AIC, bootstrapping	Yes	Yes	Gaussian, Poisson, Binomial …	Residual variance modelled as Gamma distribution	(Rönnegård et al. 2010)
R-INLA	Free	Approximate Bayesian	credible intervals for ΔV	Yes	Yes	Gaussian, Poisson, Binomial …		(Lindgren and Rue 2015)
MCMCglmm	Free	Bayesian	DIC, credible intervals for ΔV	Yes	Yes	Gaussian, Poisson, Binomial …		(Hadfield 2010)
brms	Free	Bayesian	WAIC, LOO, credible intervals for ΔV	Yes	Yes	Gaussian, Poisson, Binomial …	Residual variance modelled as log-normal distribution	(Bürkner 2017)

222   In many cases, researchers are also interested in whether the difference in variance

223   components have a biologically meaningful effect. In other words, when asking questions

224   about whether variance components vary between environments, we are mostly interested

225   in the *magnitude of the difference* in these components across environments. While model

226   comparison of LMMs can help us understand whether a statistically detectable difference is

227   observable across environments, the magnitude of the difference can only be determined

228   by examining the difference in variance components among environment: ΔV estimated as

229   V_{E2} - V_{E1} in our case. When the trait of interest is expressed as standard deviation units (i.e.

230   mean centered and scaled to the standard deviation of the dataset across all populations

231   and environments), this difference can be considered an effect size for the magnitude of the

232   difference among variance components, thus making comparisons across studies possible

233   (Royauté et al. 2015; Hamilton et al. 2017; Royauté and Dochtermann 2017). Note that ΔV

234   could also be expressed on a ratio scale (V_{E2}/V_{E1}) or on a log-additive scale ($\log(V_{E2})$ - log

235   (V_{E1})). We will return to the topic of statistical significance vs. appropriate effect sizes later

236   in the paper. For now, we simply consider ΔV on an additive scale with data expressed in

237   standard unit deviations because it allows the most straightforward interpretation and

238   functions in cases where a variance component is zero or approaching zero. ΔV can be

239   calculated from the maximum likelihood estimates in a frequentist framework but

240   calculation of the uncertainty around this estimate is not straightforward and requires

241   additional steps such as bootstrapping. In a Bayesian framework, the calculations are much

242   simpler given that the distribution of ΔV can be directly estimated by taking the difference

243   in the posterior distribution of V_{E2} - V_{E1}. The posterior mode of ΔV can then be interpreted

244   as the estimated strength of ΔV, with credible intervals representing the precision around

245   this estimate.

246   In summary, approaches based on LMM and their generalized extensions allow

247   great flexibility and are well suited to study questions related to how variation in

248   phenotypic traits varies at multiple levels of organization. In the next section, we describe

249   the performance of LMMs to detect differences in variance components.

250

251   **METHODS**

252   The simulations described below focus on interpretation in the context of behavioral

253   repeatability. However, it is worth noting again that inferences about the ability to estimate

254   and detect differences in variances generalizes to the components of the ratios described in

255   Table 1.

256   *Data simulations*

257   To compare the performance of statistical procedures for detecting differences in variance

258   components and variance ratios, we performed a series of simulations based on the

259   scenarios illustrated in Fig. 2. In these scenarios a phenotypic attribute y is measured in

260   two different environments (E1 and E2) and variation occurs among and within individuals

261   (V_I and V_W respectively). In scenarios A through C the repeatability (τ) differs by an equal

262   amount between the two environments ($\Delta\tau = 0.3$), but the underlying driver of this

263   difference is either due to a difference in the among-individual variance (A), in the within-

264   individual variance (B) or in both the among and within-individual variance (C). Note that

265 for scenario C, the total variance remains the same between environments. In scenarios D

266 and E, we explore cases where the variance ratios are equal among environment, either

267 because all variance components are equal as well (D) or in spite of differences in all

268 variance components (E) (see Table S1 for exact values for all parameters).

269   Using the R statistical environment (R Core Team 2020), we generated 500 datasets for

270 each of the following combinations:

271 • Sample size varying from 20 to 200 individuals by increments of 20 for each

272   environment (sample size was equal between the two environments)

273 • Number of repeated measures taken on each individual varying from 2 to 6

274   repeated measures by increments of 1

275 • Five different scenarios of known difference in variance ratios as described in Fig. 1

276   and Table S1.

277   Each dataset was simulated by sampling from a Gaussian distribution for the random

278 (among-individual values) and the error (within-individual) terms. This resulted in a total

279 of 125,000 datasets on which we tested three different statistical procedures to detect

280 differences in variance components and variance ratios. We provide all R code for data

281 generation and analysis in the Electronic Supplementary Materials (ESM1).
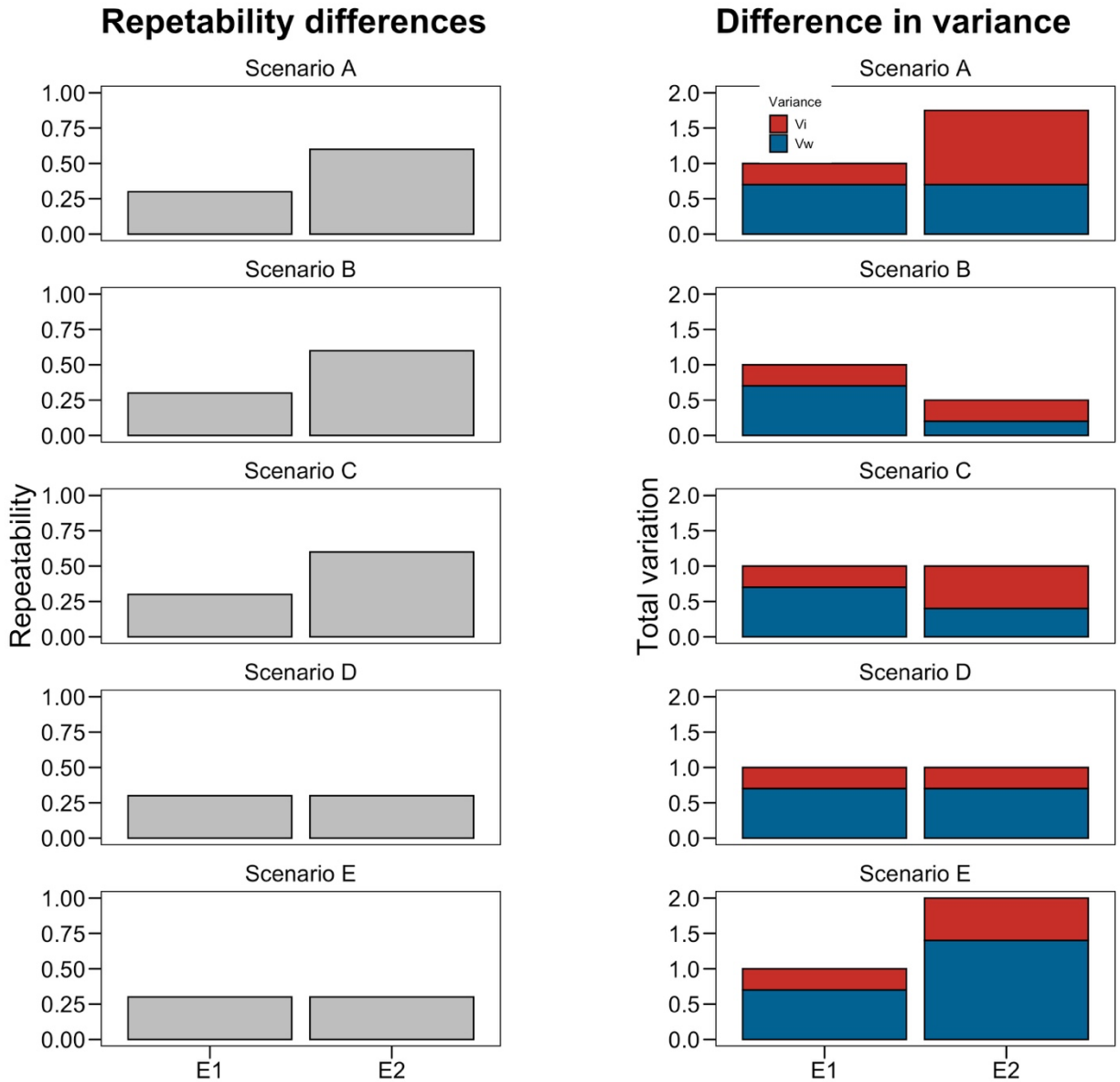
Fig. 2 Scenarios used in simulations detailing how differences or lack of difference in repeatability (right-side column) can arise from different patterns in the underlying variance components (left-side column; exact values can be found in Table S1). Scenarios A-C correspond to cases where the total variation differs between two environments (E1 and E2) due to differences in the among-individual variance (V_I, A), the within-individual variance (V_W, B) or both (C). Scenarios D-E indicate cases where the ratios remain constant across environments, because all variance components are identical (D) or in spite of variance component being different among environments (E)

292   *Comparison of confidence interval overlap from separate mixed models*

293   We first compared the overlap of 83 % confidence intervals for variance component when

294   estimated from separate linear mixed models. We specified one mixed model for

295   environment 1 and one for environment 2. These models are a simplified version of the one

296   presented in equation (3):

297    $y_{ij} = \beta_0 + ID_{0i} + e_{0ij}$ (equation 2)

298    $ind_{0i} \sim \mathcal{N}(0, V_{ID})$;

299    $e_{0ij} \sim \mathcal{N}(0, V_e)$

300   The experimental units in the environment of interest are included as random effects and

301   no additional fixed effect are needed. Upon fitting these models, we computed 83 %

302   confidence intervals for the among and within-individual variance. Datasets where these

303   intervals did not overlap were considered as statistically different.

304   *Frequentist LMM with AIC model comparison*

305   Our second approach was to fit the LMM approach described above and test for the for the

306   significance of the difference in among- and within-individual variance using likelihood

307   ratio tests. Specifically, we specified four different mixed models corresponding to the four

308   different possibilities by which variance components may differ (Royauté et al. 2019;

309   Bucklaew and Dochtermann 2021):

310   • Model 1: a null model where the among (V_I) and within-individual variance (V_W)

311   was kept constant among environments.

312   • Model 2: a model where only the among-individual variance differs among

313   environments, while the within-individual variance is kept constant ($V_I \neq$ & $V_W =$)

314   • Model 3: a model where only the within-individual variance differs among

315   environments while the among-individual variance is kept constant ($V_I =$ & $V_W \neq$)

316   • Model 4: a model where both the among and within-individual variance were

317   allowed to vary among environments ($V_I \neq$ & $V_W \neq$)

318   For each dataset combination, we then compared each model's Aikaike's Information

319   Criterion value (AIC). AIC allows the comparison of relative fit of statistical models and

320   models with lower AIC values indicate better support relative to competing models. These

321   simulations and this analytical framework are similar to previously used approaches (Shaw

322   1991; Jenkins 2011; Tüzün et al. 2017). These models were specified using the *nlme*

323   package for mixed models (Pinheiro and Bates 2006) using Restricted Maximum

324   Likelihood (REML).

325   *Bayesian LMM and difference in variance components*

326   We next fit a mixed model where variances among and within units were allowed to vary

327   between environments (as in model 4 described above) to each randomly generated

328   dataset. We calculated the posterior mode for the difference in variance components

329   (calculated as $\Delta V = V_{E2} - V_{E1}$) and estimated the 95 % credible intervals based on the

330   Highest Posterior Density of this distribution. 95 % credible intervals excluding 0 were

331   taken to indicate statistically detectable differences in variance components among

332   environments. All models were run with the *MCMCglmm* package (Hadfield 2010) using

333   default iteration settings to shorten computing time (13000 iterations, 3000 burn-in

334   iterations and thinning interval of 10 iterations). We used priors that were minimally

335   informative for the variance components (See ESM1 and ESM3 for prior specification and a
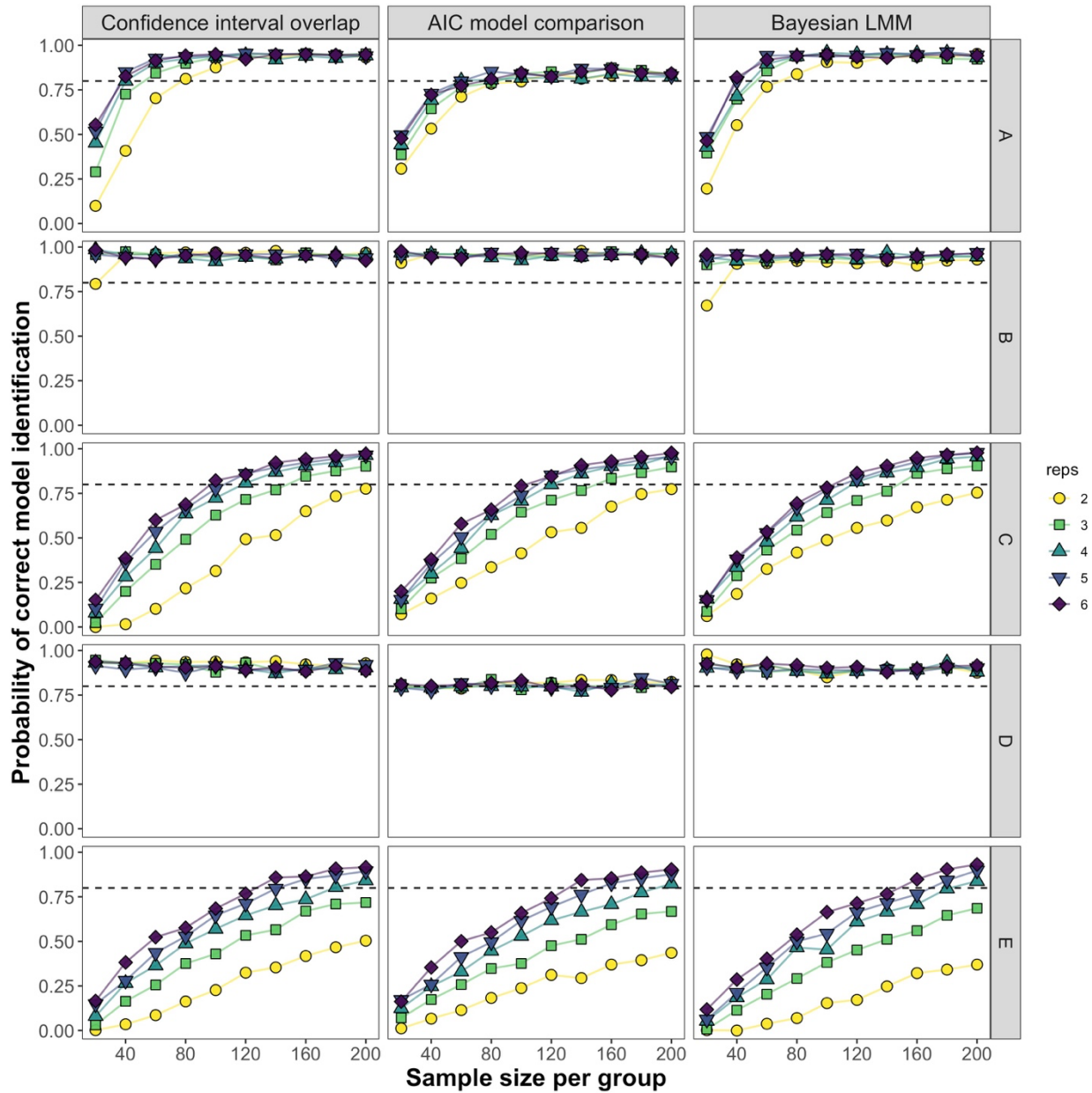
336   discussion on priors).

337   *Probability of correct model identification, precision, bias and accuracy estimations*

338   We calculated the probability of detecting the model with the correct difference in variance

339   components (hereafter "abridged" to probability of correct model identification), precision,

340   relative bias and accuracy under each scenario and sampling design to compare the

341   performance of maximum likelihood and Bayesian mixed models. For Method 1 (overlap of

342   83 % intervals), we assigned values of 1 when significant differences in variance

343   components were detected in directions predicted by the data generating process, and 0

344   otherwise. For Method 2, we calculated the probability of correct model identification as

345   the proportion of times the model with the lowest AIC matched the generating model. For

346   Method 3, we calculated whether a given model detected a difference in variance

347   components based on the overlap of the 95 % credible intervals of the ΔV posterior

348   distribution with 0. As in Method 1, we then assigned values of 0 or 1 based on whether the

349   detected difference matched with the data generation process of the corresponding

350   scenario. We calculated the probability of correct model identification as the proportion of

351   analyzed datasets in which we detected differences in the direction predicted by each

352   scenario and statistical method. Precision, indicating the similarity of the results produced

353   by simulations with a given scenario, was calculated as the difference between 25 % and 75

354   % quantiles of estimates (van de Pol 2012). To calculate the relative bias (in %) for each

355   statistical approach by scenario, we calculated the mean difference between the expected

356   value and the value observed in each of the 500 simulations. Finally, we report the root

357   mean square of error (RMSE) for each scenario and sample sizes. This metric calculates

358   how close estimates are to the expected values and serves as an estimate of the accuracy of

359   each statistical approach by scenario.

360   **RESULTS**

361   The probability of correctly detecting differences in variance components did not differ

362   substantially between frequentist and Bayesian methods of estimation (Fig. 3). The highest

363   probability of correct model identification was observed for cases where the variance ratio

364   differs as a result of changes to the within-individual variance (scenario B) or when

365   variation remained equal between environments (scenario D). The statistical power to

366   differentiate between alternative scenarios (i.e. scenarios A, C and E) was lower, especially

367   with small sample sizes and low number of repeated measures (Fig. 3). Importantly, no

368   statistical method seemed to outperform all others across scenarios. Our results are

369   consistent with previous simulations showing that the among-individual variance

370   component is particularly difficult to estimate at small sample sizes (Dingemanse and

371   Dochtermann 2013).

Fig. 3 Effect of sampling design on the probability of correct model identification by scenario type and statistical modeling approach. Each point represents the probability of detecting the correct differences in variance averaged over 500 simulated datasets for a given sample size (n: number of individuals measured in each population, reps: number of repeated measures per individuals). A represents a scenario where only the among-individual variance (V_I) varies between environments, B represents a case where the within-individual variance (V_W) varies between environments, and both among and within-individual variance vary between environments in scenario C. In scenario D, all variance components are equal while in scenario E, variance components are different but variance ratios are equal across environments. Dashed lines correspond to 80 % threshold similar to recommendations for power analyses.

384   In scenarios B and D, the correct differences among variance components were

385   identified > 80 % of the time, even at low sample sizes (Fig. 3). In all other scenarios this

386   threshold was only reached with high sample sizes and a high number of repeated

387   measures. For scenarios C and E—which correspond to cases where the variance ratio

388   differs as a result of among-individual variance (C) or when the variance ratio remains the

389   same despite changes to both among- and within-individual variance (E)—datasets with

390   only 2 repeated measures per individual never achieved a probability of identifying the

391   generating model above 0.8, even with sample sizes above 200 units per environment (i.e. a

392   minimum of 800 total measurements, Fig. 3). Increasing the number of repeated measures

393   only marginally alleviated the problem. For example, in scenario C, only datasets with 4 or

394   more repeated measures per individual reached statistical power above 0.8 with sample

395   sizes above 120 individuals per environment, which is higher than many ecological or

396   evolutionary studies can provide under realistic scenarios.

397   Note that for AIC model comparison, we calculated power as the number of times

398   the best model corresponded to the generating model. A more conservative approach is to

399   calculate the proportion of times the best model is at least 2 AIC units lower than the

400   second model. This method corresponds to a common threshold to detect statistically

401   distinct models (Burnham and Anderson 1998). When using this more conservative

402   threshold (Fig. S1), datasets generated according to scenarios A and D were never

403   statistically distinguishable from non-generating models, although the correct model was

404   consistently ranked as the best model. This discrepancy is likely because when the

405   generating model does not include differences in the within-individual variability

406   (scenarios A and D), sampling error is erroneously identified as heterogeneity. At smaller

407   sample sizes this error is greater on average, and thus detectable. At larger sample sizes

408   this sampling error is smaller but more easily detected and therefore manifests as a

409   difference between groups. To address this, in addition to measures of variance differences

410   like the described ΔV statistic, researchers should also compare mean-standardized

411   variance estimates like the coefficient of variation or Houle's evolvability between groups

412   (Houle 1992; Hansen et al. 2011; Dochtermann and Royauté 2019).

413   The comparison of relative bias, precision, and accuracy among statistical methods

414   produced mixed results. On average, Bayesian LMMs consistently underestimated the

415   among-individual variance for scenarios in which the among-individual variance differed

416   between environments (scenarios A, C, and E) resulting in a bias at small sample sizes (Fig.

417   S2). However, Bayesian LMMs also had higher precision and accuracy compared to

418   maximum likelihood (Fig. S3, S4). This means that Bayesian estimates tend to be

419   consistently more conservative than maximum likelihood regarding the magnitude of the

420   among-individual variance but that these estimates nonetheless more closely matched

421   simulation conditions.

422   **DISCUSSION**

423   Comparing variability across datasets is important for many questions in evolutionary

424   ecology (e.g. Table 1). However, variance ratios are not sufficient to address questions

425   about how variation is expressed across environments, populations, or sexes. The inability

426   to determine why groups differ based on ratios is in addition to the numerous conceptual

427   and theoretical problems inherent to the estimation of variance ratios (Houle 1992; Hansen

428   et al. 2011). Instead, many questions require the direct comparison of variances.
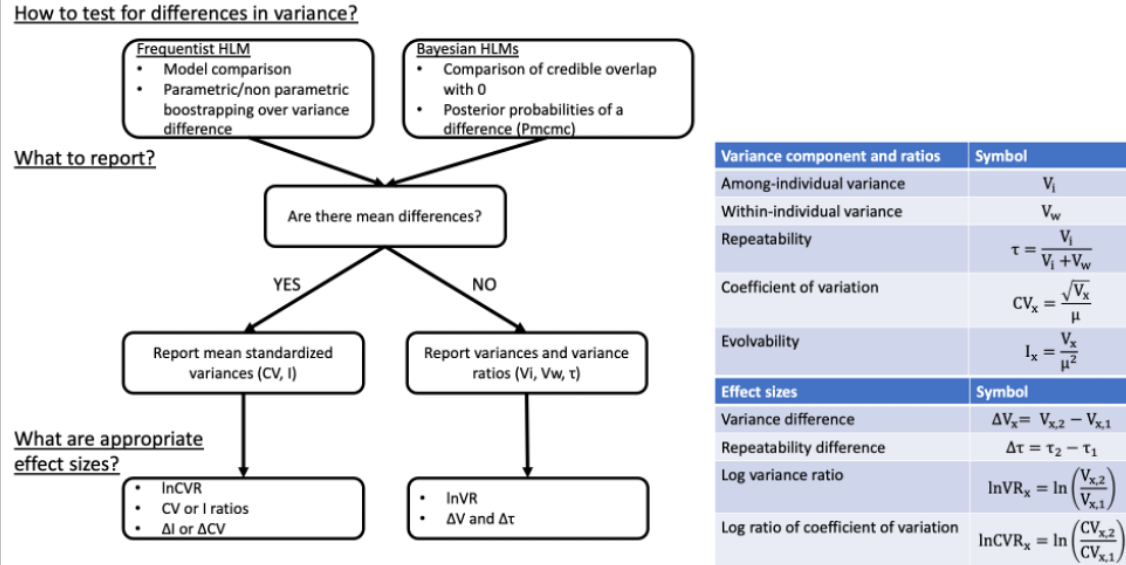
429   *What are appropriate sample sizes for detecting differences in variance?*

430   Our simulations show that regardless of the statistical methods used, comparing

431   variance components across groups is a "data hungry" question. Scenarios where the

432   among-individual variance differed between environments were particularly hard to detect

433   at low sample sizes. Note that our objective was not to provide a full exploration of

434   parameter space. Instead, we focused on a subset of scenarios that are likely to be common

435   in ecology and evolution (Fig. 2). Based on our simulations, the probability to detect

436   differences in variance components will depend in large part on the ability to estimate the

437   among-individual variance component (V_I). In the most complex case where differences

438   occur among and within-individuals (scenario E), researchers would require a minimum of

439   1,600 observations to correctly detect differences (i.e, 200 individuals measured 4 times in

440   each environment). This is far higher than sample sizes needed for single populations,

441   where moderate repeatabilities only need ~100 observations to be estimated with > 0.8

442   power (at least 25 individuals measured 4 times to detect a repeatability of 0.3; see

443   (Dingemanse and Dochtermann 2013).

444   Given these challenges, we recommend that researchers conduct power calculations

445   prior to the experiment whenever possible (see R code for *a priori* power analyses in ESM2

446   and an R Markdown tutorial in ESM3). If not, a simple rule for sampling can be to estimate

447   the sample size needed to detect the lowest among-individual variance value of interest

448   (see, for example, (Martin et al. 2011; van de Pol 2012; Dingemanse and Dochtermann

449   2013) and multiplying that sample size by the number of experimental groups involved.

450

A

How to test for differences in variance?

Frequentist HLM
- Model comparison
- Parametric/non parametric boostrapping over variance difference

Bayesian HLMs
- Comparison of credible overlap with 0
- Posterior probabilities of a difference (Pmcmc)

What to report?

Are there mean differences?

YES → Report mean standardized variances (CV, I)

NO → Report variances and variance ratios (V_i, V_w, τ)

What are appropriate effect sizes?

(YES)
- lnCVR
- CV or I ratios
- ΔI or ΔCV

(NO)
- lnVR
- ΔV and $\Delta\tau$

Variance component and ratios	Symbol
Among-individual variance	V_i
Within-individual variance	V_w
Repeatability	$\tau = \dfrac{V_i}{V_i + V_w}$
Coefficient of variation	$CV_x = \dfrac{\sqrt{V_x}}{\mu}$
Evolvability	$I_x = \dfrac{V_x}{\mu^2}$

Effect sizes	Symbol
Variance difference	$\Delta V_x = V_{x,2} - V_{x,1}$
Repeatability difference	$\Delta\tau = \tau_2 - \tau_1$
Log variance ratio	$lnVR_x = \ln\left(\dfrac{V_{x,2}}{V_{x,1}}\right)$
Log ratio of coefficient of variation	$lnCVR_x = \ln\left(\dfrac{CV_{x,2}}{CV_{x,1}}\right)$

B

Model	Comparison	k	Log-Likelihood	AIC	ΔAIC
Model 1	V_i & V_w =	4	-6538.02	13084.04	383.84
Model 2	V_i ≠ & V_w =	6	-6493.45	12998.90	298.70
Model 3	V_i = & V_w ≠	5	-6383.17	12776.34	76.14
Model 4	**V_i & V_w ≠**	**7**	**-6343.10**	**12700.20**	**0.00**

Variance and ratio	Environment	
	Low elevation Median [95% CI]	High elevation Median [95% CI]
V_i	165.96 [132.39; 199.65]	662.29 [540.31; 815.68]
V_w	38.26 [33.71; 42.64]	167.07 [148.77; 187.21]
τ	0.81 [0.77; 0.84]	0.80 [0.76; 0.84]

Difference	Δ (High − Low) Median [95% CI]	Pmcmc
V_i	**493.89 [370.25; 648.69]**	**1.00**
V_w	**128.76 [107.57; 148.4]**	**1.00**
τ	-0.01 [-0.06; 0.04]	0.69

451
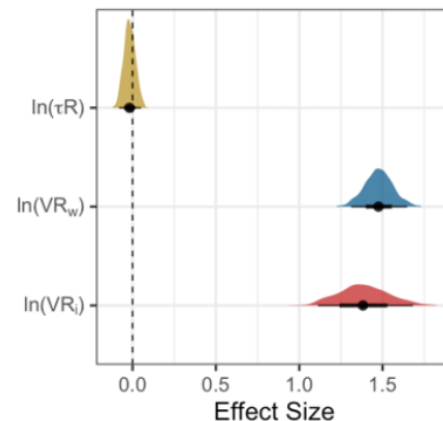
Fig. 4 A) Flowchart showing decision rules regarding how to test for differences in variance components, which metrics to report and which effect sizes can be calculated, along with their definitions in table format. B) Reporting example based on the simulated case study in Fig. 1B, C. The first Table used REML model selection with AIC to compare the support for different hypotheses for how variance components of aggression may differ between the low and high elevation populations. The best model is one where among and within-individual variances are higher in the high elevation population. The second Table compares all components by environment (posterior medians and 95 % credible intervals estimated from a Bayesian mixed model with model 4, note that frequentist confidence interval can also be reported using non-parametric bootstrapping as shown in ESM3). Finally, because aggression does not differ on average between populations, lnVR is an appropriate metric to report the effect size for the difference in variance between populations.

How to report results? Statistical significance vs. effect sizes

465   Given the issues discussed above, how should researchers interested in ecological

466   and evolutionary variation design their studies and report their findings? We suggest that

467   researchers report their results in a manner that focuses on the magnitude of the difference

468   in variability between experimental groups rather than solely focus on statistical

469   significance.

470   To this effect, we believe that reporting the results of the full model rather than just

471   the most parsimonious model will be most appropriate in most cases (i.e. model 4 in our

472   conceptual example). This is because model selection only gives information on whether

473   differences among groups are statistically detectable. In contrast, questions regarding the

474   magnitude and precision of the estimated differences are answerable only with

475   interpretation of the most complete statistical model (see tutorial in ESM3).

476   In addition to presenting results of the full model, we suggest that measures of effect

477   sizes for the differences in variance component also be presented. As reported above, ΔV

478   provides a simple metric to estimate the magnitude of these differences, but it is by no

479   mean the only one. In our theoretical example, the mean trait value did not differ by

480   environments, but in many cases mean and variance are related. In such cases, using

481   comparisons based on Houle's (1992) I^2 value or coefficients of variation for each

482   component as opposed to variance component themselves can be preferable (Hansen et al.

483   2011; Dochtermann and Royauté 2019). Effect sizes based on the coefficient of variation

484   can also be calculated within an LMM framework as described by (Nakagawa et al. 2015)

485   (see also (Carmona et al. 2016; Fontana et al. 2018) for approaches relevant to functional

486   trait diversity).

487   We provide a synthetic guide for which statistical tests and effect sizes are most

488   appropriate depending on the nature of the dataset in Fig. 4A. Returning to our dahu

489   example, an appropriate analysis of the difference in aggression variance would follow the

490   tables and figures from Fig. 4B. Here the repeatability is unchanged between environments

491   (posterior median [95 % credible interval]; $\Delta\tau$ = -0.01 [-0.06; 0.04], probability of

492   difference: Pmcmc = 0.68). However, the high-elevation population shows significantly

493   higher variation among and within-individuals (ΔVR_I = 493.89 [370.25; 648.69], Pmcmc =

494   1.00; ΔVR_W = 128.76 [107.57; 148.40], Pmcmc = 1.00). This difference is also biologically

495   relevant since the effect sizes are also > 1 ($lnVR_I$ = 1.38 [1.10, 1.66]; ΔVR_W = 1.48 [1.31,

496   1.64]). Biologically, this means that the high elevation population is composed of

497   individuals that are more distinct in behavior compared to the low elevation population.

498   While we limited our conceptual example to comparisons between two

499   environments, the LMM approach we propose is by no mean restricted to two-groups

500   comparisons. For example, Jenkins (2011) used model comparison to tease apart the

501   relative influence of sex, species and their interaction on the expression of behavioral

502   variation in kangaroo rats. Similarly, (Coblentz et al. 2017) show how model selection

503   combined with Bayesian GLMM can allow the comparison of indices of diet specialization

504   within and among species. In both cases, model selection can provide a first pass at

505   whether differences in variance components are detectable among groups, while specific

506   pairwise comparisons of effect sizes (using ΔV or other metrics) will allow discernment of

507   the most pronounced differences in variance component. Regardless of the statistical

508   approach used, we suggest it is important that researchers clearly outline the direction and,

509   when possible, magnitude of the expected effects in their predictions.

510   Finally, our conceptual examples focus exclusively on the case of "well-behaved"

511   data with normal error distributions. While these comparisons can be made with

512   generalized extensions to LMMS (i.e. GLMMs), researchers must take extra precautions

513   when calculating and comparing the within-individual variances (i.e. the residual variance).

514   Indeed, in the case of non-Gaussian data, the residual variance depends on both the link

515   function used and how the software deals with overdispersion (additive vs. multiplicative

516   overdispersion). (Nakagawa and Schielzeth 2010)) provides a very useful and extensive

517   guide explaining how the correct residual variation can be calculated.

518   **CONCLUSIONS**

519   Variance ratios are straightforward metrics to describe how various ecological and

520   evolutionary processes occur. However, comparing these ratios across studies or group can

521   be misleading if poor attention is given to the specific variance components making up

522   those ratios. More importantly, a lack of difference in these ratios does not mean that

523   variation is expressed equally among groups. Given these limitations, we advocate for

524   techniques allowing the estimation of differences in each variance components rather than

525   focusing solely on variance ratios. The statistical tools allowing comparison of trait

526   variation have become increasingly sophisticated and now allow asking very precise

527   questions. Specifically, we can now ask how trait variation is generated and how variation

528   differs among groups. However, despite the availability of these tools, researchers

529   interested in ecological and evolutionary variation must remain careful in their study

530   designs. As our simulations show, scenarios involving differences in among-individual

531   variance are particularly difficult to detect without substantial sample sizes. Finally, we

532   hope the statistical approaches and tools for power analysis presented here will allow for

533   appropriate comparisons of trait variation in ecological and evolutionary studies.

534

REFERENCES

Aguirre JD, Hine E, McGuigan K, Blows MW (2014) Comparing G: multivariate analysis of genetic variation in multiple populations. Heredity 112:21–29. https://doi.org/10.1038/hdy.2013.12

Arnold SJ, Phillips PC (1999) Hierarchical Comparison of Genetic Variance-Covariance Matrices. Ii Coastal-Inland Divergence in the Garter Snake, *Thamnophis elegans*. Evolution 53:1516–1527. https://doi.org/10.1111/j.1558-5646.1999.tb05415.x

Austin PC, Hux JE (2002) A brief note on overlapping confidence intervals. J Vasc Surg 36:194–195. https://doi.org/10.1067/mva.2002.125015

Barr DR (1969) Using confidence intervals to test hypotheses. J Qual Technol 1:256–258

Bates D, Maechler M, Bolker B, et al (2015) Package 'lme4'

Bolnick DI, Svanbäck R, Fordyce JA, Yang LH, Davis JM, Hulsey CD, Forister ML (2002) The ecology of individuals: incidence and implications of individual specialization. Am Nat 161:1–28

Brooks ME, Kristensen K, Benthem KJ van, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Mächler M, Bolker BM (2017) glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. R J 9:378–400

Bucklaew A, Dochtermann NA (2021) The effects of exposure to predators on personality and plasticity. Ethology 127:158–165. https://doi.org/10.1111/eth.13107

Bürkner P-C (2017) brms: An R package for Bayesian multilevel models using Stan. J Stat Softw 80:1–28

Burnham KP, Anderson DR (1998) Practical use of the information-theoretic approach. In: Model Selection and Inference. Springer, pp 75–117

Carmona CP, de Bello F, Mason NW, Lepš J (2016) Traits without borders: integrating functional diversity across scales. Trends Ecol Evol 31:382–394

Chartois J, Claudel C (1945) Hunting the Dahut: A French Folk Custom. J Am Folk 58:21–24. https://doi.org/10.2307/535332

Coblentz KE, Rosenblatt AE, Novak M (2017) The application of Bayesian hierarchical models to quantify individual diet specialization. Ecology 98:1535–1547. https://doi.org/10.1002/ecy.1802

Dingemanse NJ, Dochtermann NA (2013) Quantifying individual variation in behaviour: mixed-effect modelling approaches. J Anim Ecol 82:39–54

575   Dochtermann NA, Roff DA (2010) Applying a quantitative genetics framework to
576   behavioural syndrome research. Philos Trans R Soc B Biol Sci 365:4013–4020.
577   https://doi.org/10.1098/rstb.2010.0129

578   Dochtermann NA, Royauté R (2019) The mean matters: going beyond repeatability to
579   interpret behavioural variation. Anim Behav 153:147–150.
580   https://doi.org/10.1016/j.anbehav.2019.05.012

581   Fontana S, Thomas MK, Moldoveanu M, Spaak P, Pomati F (2018) Individual-level trait
582   diversity predicts phytoplankton community properties better than species richness
583   or evenness. ISME J 12:356

584   Gilmour AR, Gogel BJ, Cullis BR, Welham Sj, Thompson R (2015) ASReml user guide release
585   4.1 structural specification. Hemel Hempstead VSN Int Ltd

586   Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models:
587   the MCMCglmm R package. J Stat Softw 33:1–22

588   Hamilton JA, Royauté R, Wright JW, Hodgskiss P, Ledig FT (2017) Genetic conservation and
589   management of the California endemic, Torrey pine (*Pinus torreyana* Parry):
590   Implications of genetic rescue in a genetically depauperate species. Ecol Evol
591   7:7370–7381

592   Hansen TF, Pélabon C, Houle D (2011) Heritability is not Evolvability. Evol Biol 38:258.
593   https://doi.org/10.1007/s11692-011-9127-6

594   Hector A (2021) The New Statistics with R: An Introduction for Biologists. Oxford
595   University Press

596   Houle D (1992) Comparing evolvability and variability of quantitative traits. Genetics
597   130:195–204

598   Jacquat MS (1995) Le dahu: monographie ethno-étho-biologique publiée à l'occasion de
599   l'exposition inaugurée le 1er avril 1995. Editions de la Girafe, Musée d'histoire
600   naturelle

601   Jenkins SH (2011) Sex differences in repeatability of food-hoarding behaviour of kangaroo
602   rats. Anim Behav 81:1155–1162

603   Lessells CM, Boag PT (1987) Unrepeatable repeatabilities: a common mistake. The Auk
604   104:116–121

605   Lindgren F, Rue H (2015) Bayesian Spatial Modelling with R-INLA. J Stat Softw 63:1–25.
606   https://doi.org/10.18637/jss.v063.i19

607   MacGregor-Fors I, Payton ME (2013) Contrasting Diversity Values: Statistical Inferences
608   Based on Overlapping Confidence Intervals. PLOS ONE 8:e56794.
609   https://doi.org/10.1371/journal.pone.0056794

610   Martin JG, Nussey DH, Wilson AJ, Réale D (2011) Measuring individual differences in
611   reaction norms in field and experimental studies: a power analysis of random
612   regression models. Methods Ecol Evol 2:362–374

613   Mousseau TA, Roff DA (1987) Natural selection and the heritability of fitness components.
614   Heredity 59:181

615   Nakagawa S, Poulin R, Mengersen K, Reinhold K, Engqvist L, Lagisz M, Senior AM (2015)
616   Meta-analysis of variation: ecological and evolutionary applications and beyond.
617   Methods Ecol Evol 6:143–152. https://doi.org/10.1111/2041-210X.12309

618   Nakagawa S, Schielzeth H (2010) Repeatability for Gaussian and non-Gaussian data: a
619   practical guide for biologists. Biol Rev 85:935–956

620   Nakagawa S, Schielzeth H (2012) The mean strikes back: mean–variance relationships and
621   heteroscedasticity. Trends Ecol Evol 27:474–475.
622   https://doi.org/10.1016/j.tree.2012.04.003

623   Pinheiro J, Bates D (2006) Mixed-Effects Models in S and S-PLUS. Springer Science &
624   Business Media

625   Roff D (2002) Comparing Gmatrices: A Manova Approach. Evolution 56:1286–1291.
626   https://doi.org/10.1111/j.0014-3820.2002.tb01439.x

627   Roff DA, Prokkola JM, Krams I, Rantala MJ (2012) There is more than one way to skin a G
628   matrix. J Evol Biol 25:1113–1126. https://doi.org/10.1111/j.1420-
629   9101.2012.02500.x

630   Rönnegård L, Shen X, Alam M (2010) hglm: A Package for Fitting Hierarchical Generalized
631   Linear Models. R J 2:20–28

632   Royauté R, Buddle, CM, Vincent (2015) Under the influence: Sublethal exposure to an
633   insecticide affects personality expression in a jumping spider. 29:

634   Royauté R, Dochtermann NA (2017) When the mean no longer matters: Developmental diet
635   affects behavioral variation but not population averages in the house cricket (*Acheta
636   domesticus*). Behav Ecol 28:337–345

637   Royauté R, Garrison C, Dalos J, Berdal MA, Dochtermann NA (2019) Current energy state
638   interacts with the developmental environment to influence behavioural plasticity.
639   Anim Behav 148:39–51

640 Santostefano F, Wilson AJ, Araya-Ajoy YG, Dingemanse NJ (2016) Interacting with the
641   enemy: indirect effects of personality on conspecific aggression in crickets. Behav
642   Ecol 27:1235–1246

643 Schenker N, Gentleman JF (2001) On Judging the Significance of Differences by Examining
644   the Overlap Between Confidence Intervals. Am Stat 55:182–186.
645   https://doi.org/10.1198/000313001317097960

646 Shaw RG (1991) The Comparison of Quantitative Genetic Parameters Between Populations.
647   Evolution 45:143–151. https://doi.org/10.1111/j.1558-5646.1991.tb05273.x

648 Tüzün R N, Müller, S, Koch, K, Stoks (2017) Pesticide-induced changes in personality
649   depend on the urbanization level. 134:

650 van de Pol M (2012) Quantifying individual variation in reaction norms: how study design
651   affects the accuracy, precision and power of random regression models. Methods
652   Ecol Evol 3:268–280

653 Violle C, Enquist BJ, McGill BJ, Jiang LIN, Albert CH, Hulshof C, Jung V, Messier J (2012) The
654   return of the variance: intraspecific variability in community ecology. Trends Ecol
655   Evol 27:244–252

656 White SJ, Pascall DJ, Wilson AJ (2020) Towards a comparative approach to the structure of
657   animal personality variation. Behav Ecol 31:340–351.
658   https://doi.org/10.1093/beheco/arz198

659 Wilson AJ (2018) How should we interpret estimates of individual repeatability? Evol Lett
660   2:4–8. https://doi.org/10.1002/evl3.40

661 Wilson AJ, Réale D, Clements MN, Morrissey MM, Postma E, Walling CA, Kruuk LEB, Nussey
662   DH (2010) An ecologist's guide to the animal model. J Anim Ecol 79:13–26.
663   https://doi.org/10.1111/j.1365-2656.2009.01639.x

664

Supporting Information (SI) and Electronic Supplementary Materials (ESM)

ESM 1 Raw data from simulations along with R code for data analysis and figures

(https://osf.io/5aw42/)

ESM 2 R code for conducting *a priori* power analysis (https://osf.io/5aw42/)

ESM 3 R tutorial for comparing variance components using *nlme*, *MCMCglmm* and *brms*

packages (https://osf.io/5aw42/)

Table S1 Scenarios tested in simulations to estimate the power to detect differences in

variance components of varying magnitude

Fig. S1 Effect of sampling design on the probability to detect differences in variance

components by scenario type and statistical modeling approach with ΔAIC > 2 threshold for

model comparison

Fig. S2 Effect of sampling design on relative bias by scenario type and statistical modeling

approach

Fig. S3 Effect of sampling design on estimate precision (width of the interquartile interval)

by scenario type and statistical modeling approach

Fig. S4 Effect of sampling design on model accuracy (estimated as the root mean square of

error, RMSE) by scenario type and statistical modeling approach