Comparing ecological and evolutionary variability within datasets

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

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

[b] Current address: Movement Ecology Group, Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Frankfurt, Germany

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

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

Running Head: Comparing variation within datasets

This is a pre-print of an article published in *Behavioral Ecology and Sociobiology*. The final authenticated version is available online at: https://doi.org/10.1007/s00265-021-03068-3

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

Many critical questions in ecology and evolution use variance ratios, such as repeatability, heritability, or individual resource specialization, to make inferences about ecological and evolutionary processes. In many cases these inferences rely on the comparison of variance ratios among datasets (populations, sexes, or environments). In this article, we show that current approaches of drawing inferences about group differences from comparisons of ratios are inappropriate because ratios can differ due to differences in the numerator, denominator, or both. We investigated how questions regarding differences in variance ratios and constituent variance components can be evaluated using Linear Mixed Model approaches (LMMs) and provide guidance for appropriate sampling schemes under different scenarios and discuss common pitfalls associated with estimation of differences in variance component among datasets.

Running Head: Comparing variation within datasets

Keywords: Repeatability, animal personality, individual variation, mixed models, individual niche specialization, functional traits

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.

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*	$T_{\mathrm{IP/IC}} = V_{IP} \,/\, V_{IC}$	The proportion of variation attributable to within-population variance (V_{IP}) relative to the community variance (V_{IC})	(Violle et al. 2012)
		$T_{\mathrm{IC/IR}} = V_{IC} \,/\, V_{IR}$	The proportion of variation attributable to community variance (V_{IC}) relative to the regional pool variance (V_{IR})	(Violle et al. 2012)

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

To further illustrate the inferential limits of variance ratios, consider the following scenario: researchers are studying the behaviors and dietary habits of two populations of the mythical Dahu (*Dahu desterus*; Fig. 1A) at different elevations. These elusive creatures have shorter hind-legs on their left side, thus only allowing for clockwise movement (Chartois and Claudel 1945; Jacquat 1995). While measuring aggressive interactions, researchers find no differences in means between populations and similar behavioral repeatabilities (τ = 0.8; Fig. 1B). Put another way, the same relative amount of variation is attributable to individuals in each population. The researchers notice, however, that there are large differences in the among- and within-individual variances of each population. Had researchers only examined repeatabilities and mean differences they would inappropriately conclude that the populations are behaviorally equivalent. Instead, the actual variance estimates reveal that individuals from the high-altitude population are very distinct from one another in their aggressive tendencies while, at low-altitude, individuals show little departure from the population average (Fig. 1B, C).

These researchers are also curious as to whether the harsher climate at the top of the mountain range leads to a narrower dietary breadth. Researchers predict that individual resource specialization will be higher in the low elevation population, as *D. desterus* have more food options to choose from. To the researcher’s surprise, they find much higher individual resource specialization in the high-altitude population: S_1 = 0.2, S_2 = 0.8. Upon examining the specific values of among- and within-individual variation in niche, they find that these differences are a result of the high elevation population having a much narrower total niche width (Fig. 1D) while the within-individual variation in niche preference is equal between populations.

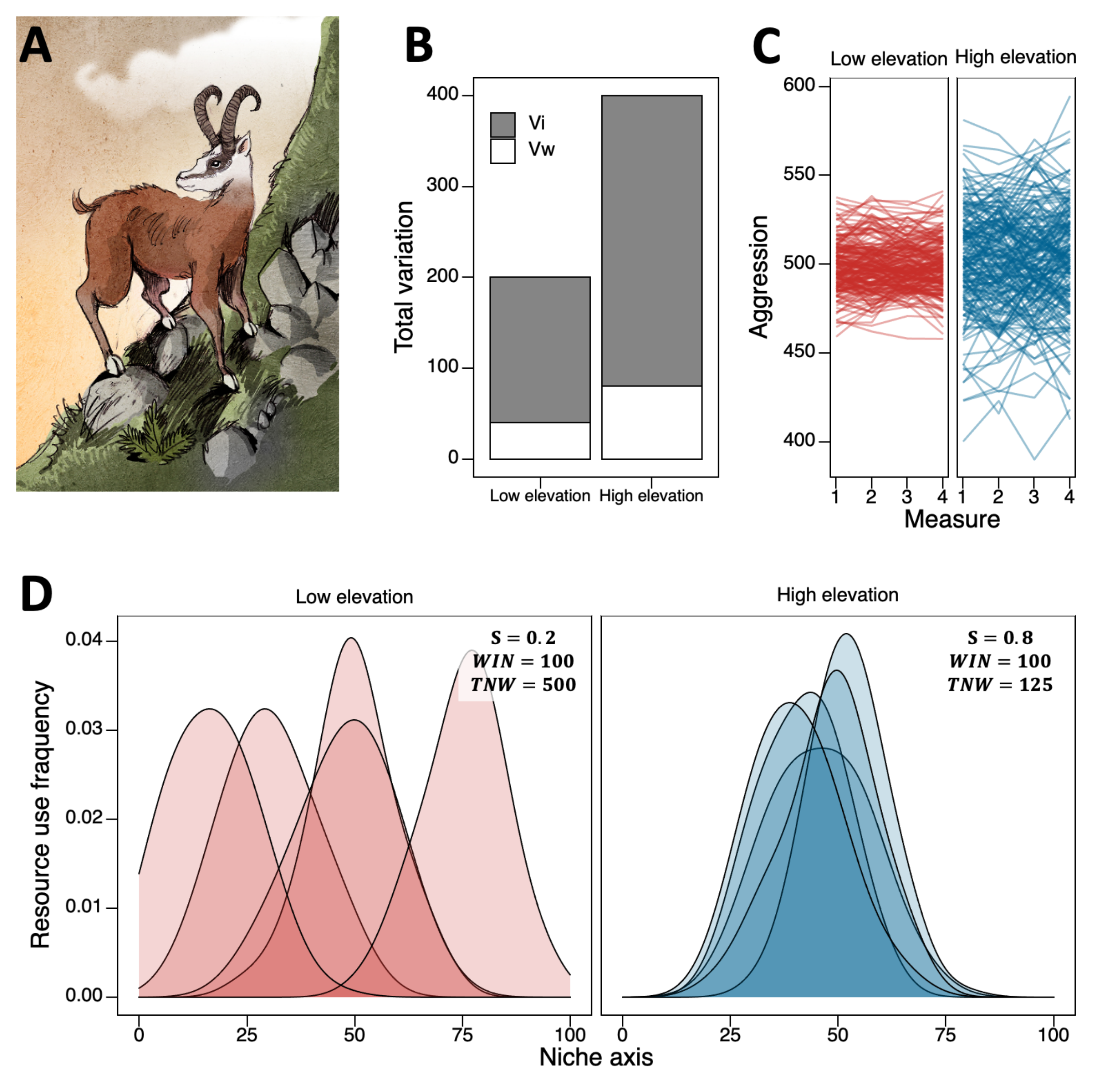


Fig. 1 Reliance on variance ratios can lead to misleading inferences. (A) The elusive Dahu (*Dahu dexterus*) in its natural environment. (B) Two populations of Dahus living at different elevations do not differ in their repeatability of aggressive interactions (τ). (C) By plotting the individual aggression 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)

This means that it is the difference in diet preference among individuals that drives the difference between the two populations. With more varied resources available at low elevation, each individual can specialize along the total niche axis, yet the breadth of diet preference within-individuals is the same between populations.

For both traits, exclusive reliance on ratios would have led to either inappropriate or incomplete inferences (i.e. inappropriately concluding behavioral equivalence and incompletely recognizing the basis of differences in apparent specialization). Due to these problems with interpretations of variance ratios (Houle 1992; Dochtermann and Royauté 2019), what would be of greater use to researchers is to instead evaluate differences in specific variance components.

A statistical framework for comparing variance components

The statistical procedures necessary for the estimation of variance components and ratios within a single population have been the subject of much attention (e.g. mixed models for repeatability: Dingemanse and Dochtermann 2013; animal models for heritability: Wilson et al. 2010; individual niche specialization:Bolnick et al. 2002; Coblentz et al. 2017; functional trait variation: Nakagawa and Schielzeth 2012; Violle et al. 2012; Carmona et al. 2016). There is also a long history in quantitative genetics regarding the comparison of variances and *co*variance structures among groups (Shaw 1991; Arnold and Phillips 1999; Roff 2002; Roff et al. 2012; Aguirre et al. 2014). Unfortunately, these quantitative genetic approaches have been poorly disseminated across fields (but see (Dochtermann and Roff 2010; White et al. 2020). Here we describe and investigate methods for detecting differences in variance components amongst groups. Specifically, we compare the strength and weaknesses of three statistical approaches: comparison of confidence intervals, model comparison with AIC, and Bayesian estimation of the difference in variance components. While this selection of methods encompasses very different philosophical approaches to data analysis, all three are routinely used in the estimation of repeatability and other ratios.

We consider a scenario where a phenotypic attribute, y, is measured repeatedly for individual organisms occupying one of two different environments (E1 and E2) and in which variation occurs among and within-individuals (V_I and V_W respectively). In the following sections we focus on differences in individual variation and repeatability. Note, however, that this scenario can also be expanded to the comparison of diet specialization for individuals occupying different environments or how functional traits vary among and within species in two different environments.

An easy way to compare these variance components and their ratios ($\tau = V_I/(V_I + V_W)$) is to estimate the variance components for each environment in separate statistical models. We can then test for differences in variances and ratios by environment based on whether estimate confidence intervals overlap or not. While straightforward, this method suffers from two key limitations. First, basing inference on the overlap of 95 % confidence intervals is overly conservative (Barr 1969), especially when sample size is low. It is instead whether the confidence interval for the *difference* in variances excludes 0 that is relevant for drawing inferences. This difference cannot be directly estimated from the approach we have described. However, statistical significance can still be assessed by comparing the overlap of the 83% confidence intervals for variance components, a threshold that provides a better approximation for an $\alpha = 0.05$ for the null hypothesis of no difference (Schenker and Gentleman 2001; Austin and Hux 2002; MacGregor-Fors and

Payton 2013; Hector 2021). Second, by estimating variance components in separate statistical models, the hierarchical structure of the data, i.e. the variance components nested within the environments, has been broken. As a result, potential average differences in the traits of interest are not appropriately tested.

Instead, we suggest that a more appropriate procedure would be the use of a Linear Mixed Model (LMM) where the among- and within-individual variance is estimated for each environment within the same statistical model. This statistical model can be described by the following equation:

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

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

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

where y_{ij} describes the phenotypic traits for the *i*th individual and *j*th observation. ID_{0i}, is the deviation from an overall intercept, β_0, for the *i*th individual. β_1 represents the regression coefficient for the fixed effect of environment (here a contrast coefficient). The random intercepts and residual variance (e_{0ij}) both follow a multivariate normal distribution, and Ω_{ID} and Ω_e, are the variance-covariance matrices at the among- and within-individual levels respectively.

The diagonal elements of these matrices represent the among- and within-individual variances in each environment: E_1 and E_2. The off-diagonal elements represent the cross-environment correlation (set to 0 if individuals are only ever evaluated in one of the two environments). This formulation has the advantage of allowing considerable flexibility in the specification of the statistical models considered (Dingemanse and

Dochtermann 2013). LMMs are now available for most statistical software and their generalized extensions can accommodate non-normal error distributions (Table 2).

Upon fitting LMMs, several methods are then available to determine whether a variance ratio or components of the ratio differ by environment. Specific hypotheses of which variance component differs across environment can be easily tested via model comparison. For example, a model where only the among-individual variance differs by environment can be compared to a null model where the among and within- individual variance are kept constant across developmental environments (Royauté et al. 2019). These models can be estimated within a frequentist framework via restricted maximum likelihood or a Bayesian framework and suitable decision criteria can be used to determine which model best fits the data. In the case of restricted maximum likelihood estimation, it is also possible to use likelihood ratio tests to compare these models. Note however that the proper degrees of freedom to apply to each model is unclear and additional care should be taken when using this method (Pinheiro and Bates 2006; Santostefano et al. 2016). We recommend calculating these degrees of freedom by considering each variance component as a full parameter for more conservative testing (see also the tutorial in ESM3).

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

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)

In many cases, researchers are also interested in whether the difference in variance components have a biologically meaningful effect. In other words, when asking questions about whether variance components vary between environments, we are mostly interested in the *magnitude of the difference* in these components across environments. While model comparison of LMMs can help us understand whether a statistically detectable difference is observable across environments, the magnitude of the difference can only be determined by examining the difference in variance components among environment: ΔV estimated as V_{E2} - V_{E1} in our case. When the trait of interest is expressed as standard deviation units (i.e. mean centered and scaled to the standard deviation of the dataset across all populations and environments), this difference can be considered an effect size for the magnitude of the difference among variance components, thus making comparisons across studies possible (Royauté et al. 2015; Hamilton et al. 2017; Royauté and Dochtermann 2017). Note that ΔV could also be expressed on a ratio scale (V_{E2}/V_{E1}) or on a log-additive scale (log(V_{E2}) - log (V_{E1})). We will return to the topic of statistical significance vs. appropriate effect sizes later in the paper. For now, we simply consider ΔV on an additive scale with data expressed in standard unit deviations because it allows the most straightforward interpretation and functions in cases where a variance component is zero or approaching zero. ΔV can be calculated from the maximum likelihood estimates in a frequentist framework but calculation of the uncertainty around this estimate is not straightforward and requires additional steps such as bootstrapping. In a Bayesian framework, the calculations are much simpler given that the distribution of ΔV can be directly estimated by taking the difference in the posterior distribution of V_{E2} - V_{E1}. The posterior mode of ΔV can then be interpreted as the estimated strength of ΔV, with credible intervals representing the precision around this estimate.

In summary, approaches based on LMM and their generalized extensions allow great flexibility and are well suited to study questions related to how variation in phenotypic traits varies at multiple levels of organization. In the next section, we describe the performance of LMMs to detect differences in variance components.

METHODS

The simulations described below focus on interpretation in the context of behavioral repeatability. However, it is worth noting again that inferences about the ability to estimate and detect differences in variances generalizes to the components of the ratios described in Table 1.

Data simulations

To compare the performance of statistical procedures for detecting differences in variance components and variance ratios, we performed a series of simulations based on the scenarios illustrated in Fig. 2. In these scenarios a phenotypic attribute y is measured in two different environments (E1 and E2) and variation occurs among and within individuals (V_I and V_W respectively). In scenarios A through C the repeatability (τ) differs by an equal amount between the two environments ($\Delta\tau = 0.3$), but the underlying driver of this difference is either due to a difference in the among-individual variance (A), in the within-individual variance (B) or in both the among and within-individual variance (C). Note that for scenario C, the total variance remains the same between environments. In scenarios D and E, we explore cases where the variance ratios are equal among environment, either because all variance components are equal as well (D) or in spite of differences in all variance components (E) (see Table S1 for exact values for all parameters).

Using the R statistical environment (R Core Team 2020), we generated 500 datasets for each of the following combinations:

- Sample size varying from 20 to 200 individuals by increments of 20 for each environment (sample size was equal between the two environments)
- Number of repeated measures taken on each individual varying from 2 to 6 repeated measures by increments of 1
- Five different scenarios of known difference in variance ratios as described in Fig. 1 and Table S1.

Each dataset was simulated by sampling from a Gaussian distribution for the random (among-individual values) and the error (within-individual) terms. This resulted in a total of 125,000 datasets on which we tested three different statistical procedures to detect differences in variance components and variance ratios. We provide all R code for data generation and analysis in the Electronic Supplementary Materials (ESM1).

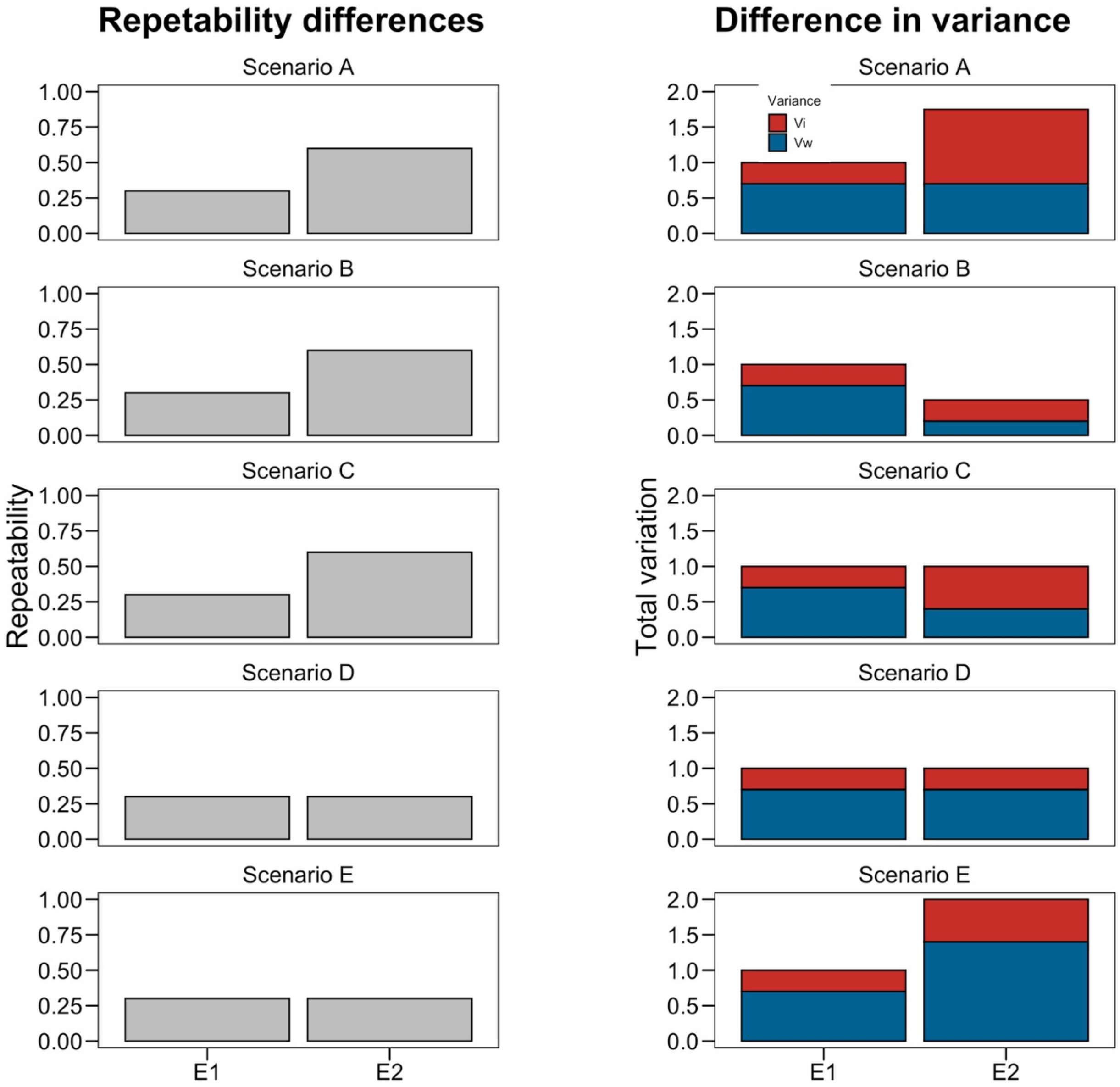


Fig. 2 Scenarios used in simulations detailing how differences or lack of difference in repeatability (right-side column) can arise from different patterns in the underlying variance components (left-side column; exact values can be found in Table S1). Scenarios A-C correspond to cases where the total 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)

Comparison of confidence interval overlap from separate mixed models

We first compared the overlap of 83 % confidence intervals for variance component when estimated from separate linear mixed models. We specified one mixed model for environment 1 and one for environment 2. These models are a simplified version of the one presented in equation (3):

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

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

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

The experimental units in the environment of interest are included as random effects and no additional fixed effect are needed. Upon fitting these models, we computed 83 % confidence intervals for the among and within-individual variance. Datasets where these intervals did not overlap were considered as statistically different.

Frequentist LMM with AIC model comparison

Our second approach was to fit the LMM approach described above and test for the for the significance of the difference in among- and within-individual variance using likelihood ratio tests. Specifically, we specified four different mixed models corresponding to the four different possibilities by which variance components may differ (Royauté et al. 2019; Bucklaew and Dochtermann 2021):

- Model 1: a null model where the among (V_I) and within-individual variance (V_W) was kept constant among environments.

- Model 2: a model where only the among-individual variance differs among environments, while the within-individual variance is kept constant (V_I ≠ & V_W =)
- Model 3: a model where only the within-individual variance differs among environments while the among-individual variance is kept constant (V_I = & V_W ≠)
- Model 4: a model where both the among and within-individual variance were allowed to vary among environments (V_I ≠ & V_W ≠)

For each dataset combination, we then compared each model's Aikaike's Information Criterion value (AIC). AIC allows the comparison of relative fit of statistical models and models with lower AIC values indicate better support relative to competing models. These simulations and this analytical framework are similar to previously used approaches (Shaw 1991; Jenkins 2011; Tüzün et al. 2017). These models were specified using the *nlme* package for mixed models (Pinheiro and Bates 2006) using Restricted Maximum Likelihood (REML).

Bayesian LMM and difference in variance components

We next fit a mixed model where variances among and within units were allowed to vary between environments (as in model 4 described above) to each randomly generated dataset. We calculated the posterior mode for the difference in variance components (calculated as $\Delta V = V_{E2} - V_{E1}$) and estimated the 95 % credible intervals based on the Highest Posterior Density of this distribution. 95 % credible intervals excluding 0 were taken to indicate statistically detectable differences in variance components among environments. All models were run with the *MCMCglmm* package (Hadfield 2010) using default iteration settings to shorten computing time (13000 iterations, 3000 burn-in iterations and thinning interval of 10 iterations). We used priors that were minimally informative for the variance components (See ESM1 and ESM3 for prior specification and a discussion on priors).

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

We calculated the probability of detecting the model with the correct difference in variance components (hereafter "abridged" to probability of correct model identification), precision, relative bias and accuracy under each scenario and sampling design to compare the performance of maximum likelihood and Bayesian mixed models. For Method 1 (overlap of 83 % intervals), we assigned values of 1 when significant differences in variance components were detected in directions predicted by the data generating process, and 0 otherwise. For Method 2, we calculated the probability of correct model identification as the proportion of times the model with the lowest AIC matched the generating model. For Method 3, we calculated whether a given model detected a difference in variance components based on the overlap of the 95 % credible intervals of the ΔV posterior distribution with 0. As in Method 1, we then assigned values of 0 or 1 based on whether the detected difference matched with the data generation process of the corresponding scenario. We calculated the probability of correct model identification as the proportion of analyzed datasets in which we detected differences in the direction predicted by each scenario and statistical method. Precision, indicating the similarity of the results produced by simulations with a given scenario, was calculated as the difference between 25 % and 75 % quantiles of estimates (van de Pol 2012). To calculate the relative bias (in %) for each statistical approach by scenario, we calculated the mean difference between the expected value and the value observed in each of the 500 simulations. Finally, we report the root mean square of error (RMSE) for each scenario and sample sizes. This metric calculates how close estimates are to the expected values and serves as an estimate of the accuracy of each statistical approach by scenario.

RESULTS

The probability of correctly detecting differences in variance components did not differ substantially between frequentist and Bayesian methods of estimation (Fig. 3). The highest probability of correct model identification was observed for cases where the variance ratio differs as a result of changes to the within-individual variance (scenario B) or when variation remained equal between environments (scenario D). The statistical power to differentiate between alternative scenarios (i.e. scenarios A, C and E) was lower, especially with small sample sizes and low number of repeated measures (Fig. 3). Importantly, no statistical method seemed to outperform all others across scenarios. Our results are consistent with previous simulations showing that the among-individual variance component is particularly difficult to estimate at small sample sizes (Dingemanse and Dochtermann 2013).

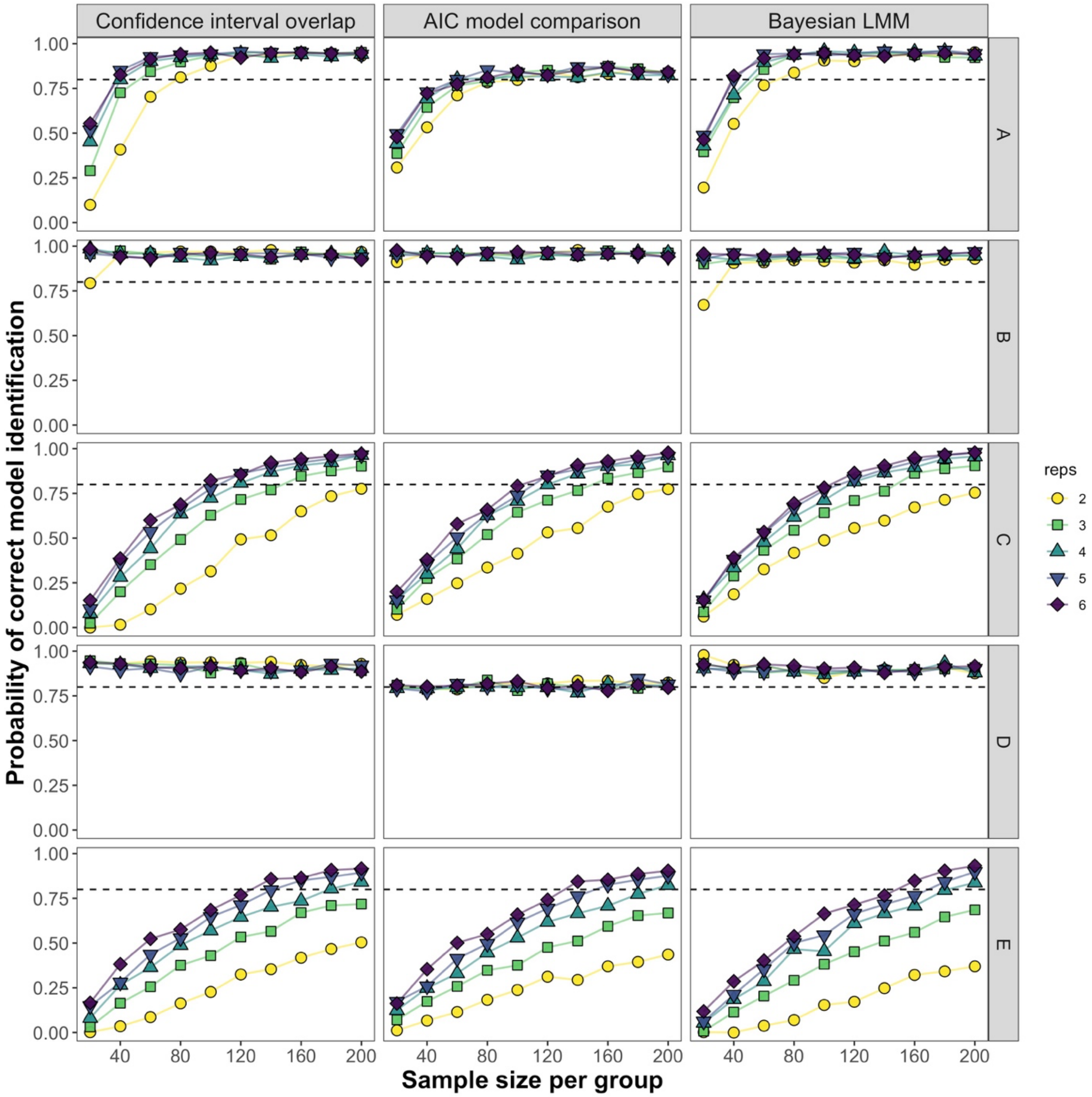


Fig. 3 Effect of sampling design on the probability of correct model identification by scenario type and statistical modeling approach. Each point represents the probability of detecting the correct differences in variance averaged over 500 simulated datasets for a given sample size (n: number of individuals 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.

In scenarios B and D, the correct differences among variance components were identified > 80 % of the time, even at low sample sizes (Fig. 3). In all other scenarios this threshold was only reached with high sample sizes and a high number of repeated measures. For scenarios C and E—which correspond to cases where the variance ratio differs as a result of among-individual variance (C) or when the variance ratio remains the same despite changes to both among- and within-individual variance (E)—datasets with only 2 repeated measures per individual never achieved a probability of identifying the generating model above 0.8, even with sample sizes above 200 units per environment (i.e. a minimum of 800 total measurements, Fig. 3). Increasing the number of repeated measures only marginally alleviated the problem. For example, in scenario C, only datasets with 4 or more repeated measures per individual reached statistical power above 0.8 with sample sizes above 120 individuals per environment, which is higher than many ecological or evolutionary studies can provide under realistic scenarios.

Note that for AIC model comparison, we calculated power as the number of times the best model corresponded to the generating model. A more conservative approach is to calculate the proportion of times the best model is at least 2 AIC units lower than the second model. This method corresponds to a common threshold to detect statistically distinct models (Burnham and Anderson 1998). When using this more conservative threshold (Fig. S1), datasets generated according to scenarios A and D were never statistically distinguishable from non-generating models, although the correct model was consistently ranked as the best model. This discrepancy is likely because when the generating model does not include differences in the within-individual variability (scenarios A and D), sampling error is erroneously identified as heterogeneity. At smaller sample sizes this error is greater on average, and thus detectable. At larger sample sizes this sampling error is smaller but more easily detected and therefore manifests as a difference between groups. To address this, in addition to measures of variance differences like the described ΔV statistic, researchers should also compare mean-standardized variance estimates like the coefficient of variation or Houle's evolvability between groups (Houle 1992; Hansen et al. 2011; Dochtermann and Royauté 2019).

The comparison of relative bias, precision, and accuracy among statistical methods produced mixed results. On average, Bayesian LMMs consistently underestimated the among-individual variance for scenarios in which the among-individual variance differed between environments (scenarios A, C, and E) resulting in a bias at small sample sizes (Fig. S2). However, Bayesian LMMs also had higher precision and accuracy compared to maximum likelihood (Fig. S3, S4). This means that Bayesian estimates tend to be consistently more conservative than maximum likelihood regarding the magnitude of the among-individual variance but that these estimates nonetheless more closely matched simulation conditions.

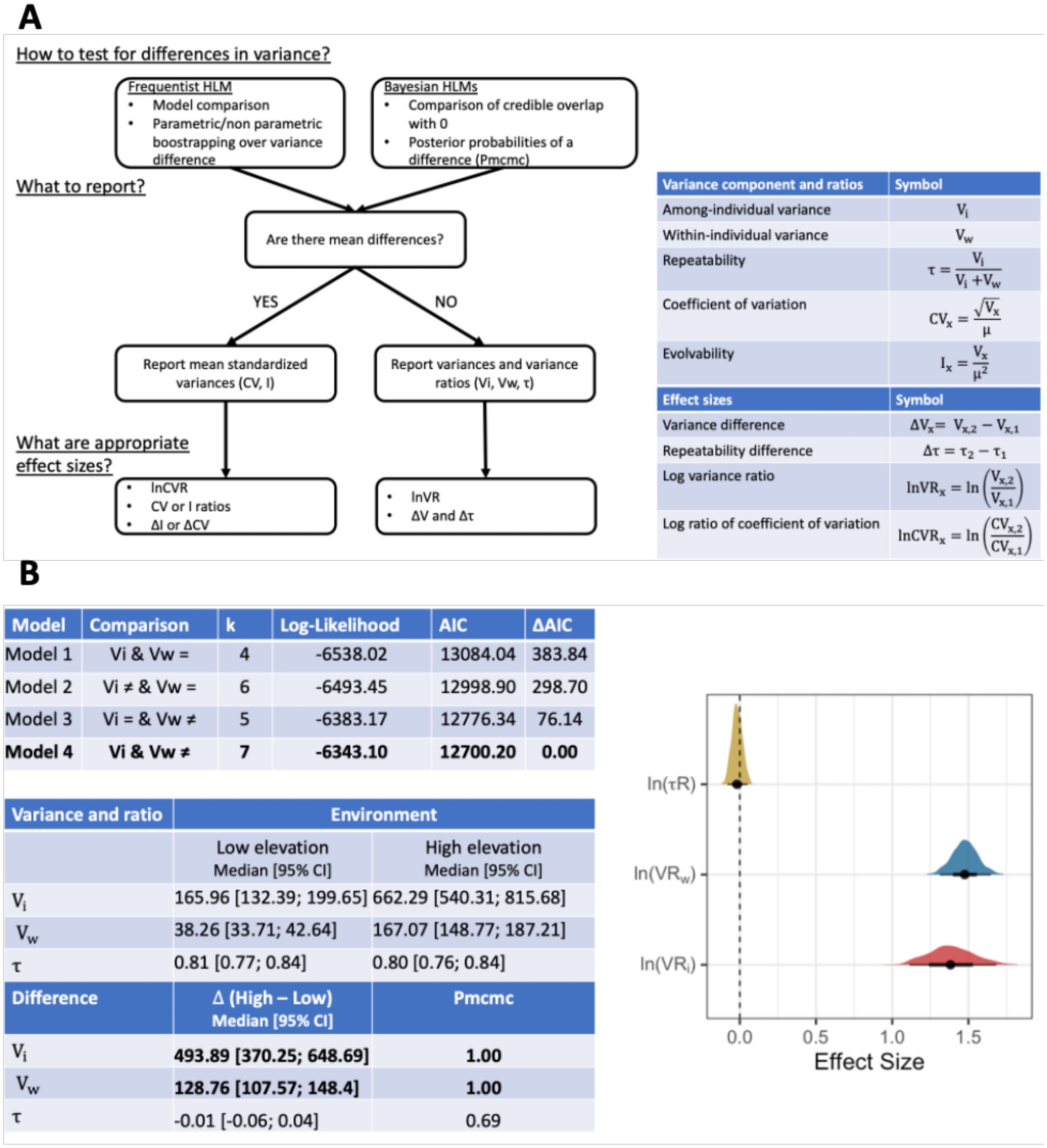
DISCUSSION

Comparing variability across datasets is important for many questions in evolutionary ecology (e.g. Table 1). However, variance ratios are not sufficient to address questions about how variation is expressed across environments, populations, or sexes. The inability to determine why groups differ based on ratios is in addition to the numerous conceptual and theoretical problems inherent to the estimation of variance ratios (Houle 1992; Hansen et al. 2011). Instead, many questions require the direct comparison of variances.

What are appropriate sample sizes for detecting differences in variance?

Our simulations show that regardless of the statistical methods used, comparing variance components across groups is a "data hungry" question. Scenarios where the among-individual variance differed between environments were particularly hard to detect at low sample sizes. Note that our objective was not to provide a full exploration of parameter space. Instead, we focused on a subset of scenarios that are likely to be common in ecology and evolution (Fig. 2). Based on our simulations, the probability to detect differences in variance components will depend in large part on the ability to estimate the among-individual variance component (V_I). In the most complex case where differences occur among and within-individuals (scenario E), researchers would require a minimum of 1,600 observations to correctly detect differences (i.e, 200 individuals measured 4 times in each environment). This is far higher than sample sizes needed for single populations, where moderate repeatabilities only need ~100 observations to be estimated with > 0.8 power (at least 25 individuals measured 4 times to detect a repeatability of 0.3; see (Dingemanse and Dochtermann 2013).

Given these challenges, we recommend that researchers conduct power calculations prior to the experiment whenever possible (see R code for *a priori* power analyses in ESM2 and an R Markdown tutorial in ESM3). If not, a simple rule for sampling can be to estimate the sample size needed to detect the lowest among-individual variance value of interest (see, for example, (Martin et al. 2011; van de Pol 2012; Dingemanse and Dochtermann 2013) and multiplying that sample size by the number of experimental groups involved.


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
NO
Report mean standardized variances (CV, I)
Report variances and variance ratios (Vi, Vw, τ)
What are appropriate effect sizes?
• lnCVR
• CV or I ratios
• ΔI or ΔCV
• lnVR
• ΔV and Δτ
B
ln(τR)
ln(VR_w)
ln(VR_i)
0.0
0.5
1.0
1.5
Effect Size

Variance component and ratios	Symbol
Among-individual variance	V_i
Within-individual variance	V_w
Repeatability	$\tau = \frac{V_i}{V_i + V_w}$
Coefficient of variation	$CV_x = \frac{\sqrt{V_x}}{\mu}$
Evolvability	$I_x = \frac{V_x}{\mu^2}$
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	$\ln VR_x = \ln\left(\frac{V_{x,2}}{V_{x,1}}\right)$
Log ratio of coefficient of variation	$\ln CVR_x = \ln\left(\frac{CV_{x,2}}{CV_{x,1}}\right)$

Model	Comparison	k	Log-Likelihood	AIC	ΔAIC
Model 1	Vi & Vw =	4	-6538.02	13084.04	383.84
Model 2	Vi ≠ & Vw =	6	-6493.45	12998.90	298.70
Model 3	Vi = & Vw ≠	5	-6383.17	12776.34	76.14
Model 4	**Vi & Vw ≠**	**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

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

Given the issues discussed above, how should researchers interested in ecological and evolutionary variation design their studies and report their findings? We suggest that researchers report their results in a manner that focuses on the magnitude of the difference in variability between experimental groups rather than solely focus on statistical significance.

To this effect, we believe that reporting the results of the full model rather than just the most parsimonious model will be most appropriate in most cases (i.e. model 4 in our conceptual example). This is because model selection only gives information on whether differences among groups are statistically detectable. In contrast, questions regarding the magnitude and precision of the estimated differences are answerable only with interpretation of the most complete statistical model (see tutorial in ESM3).

In addition to presenting results of the full model, we suggest that measures of effect sizes for the differences in variance component also be presented. As reported above, ΔV provides a simple metric to estimate the magnitude of these differences, but it is by no mean the only one. In our theoretical example, the mean trait value did not differ by environments, but in many cases mean and variance are related. In such cases, using comparisons based on Houle's (1992) I^2 value or coefficients of variation for each component as opposed to variance component themselves can be preferable (Hansen et al. 2011; Dochtermann and Royauté 2019). Effect sizes based on the coefficient of variation can also be calculated within an LMM framework as described by (Nakagawa et al. 2015) (see also (Carmona et al. 2016; Fontana et al. 2018) for approaches relevant to functional trait diversity).

We provide a synthetic guide for which statistical tests and effect sizes are most appropriate depending on the nature of the dataset in Fig. 4A. Returning to our dahu example, an appropriate analysis of the difference in aggression variance would follow the tables and figures from Fig. 4B. Here the repeatability is unchanged between environments (posterior median [95 % credible interval]; $\Delta\tau$ = -0.01 [-0.06; 0.04], probability of difference: Pmcmc = 0.68). However, the high-elevation population shows significantly higher variation among and within-individuals (ΔVR_I = 493.89 [370.25; 648.69], Pmcmc = 1.00; ΔVR_W = 128.76 [107.57; 148.40], Pmcmc = 1.00). This difference is also biologically relevant since the effect sizes are also > 1 ($lnVR_I$ = 1.38 [1.10, 1.66]; ΔVR_W = 1.48 [1.31, 1.64]). Biologically, this means that the high elevation population is composed of individuals that are more distinct in behavior compared to the low elevation population.

While we limited our conceptual example to comparisons between two environments, the LMM approach we propose is by no mean restricted to two-groups comparisons. For example, Jenkins (2011) used model comparison to tease apart the relative influence of sex, species and their interaction on the expression of behavioral variation in kangaroo rats. Similarly, (Coblentz et al. 2017) show how model selection combined with Bayesian GLMM can allow the comparison of indices of diet specialization within and among species. In both cases, model selection can provide a first pass at whether differences in variance components are detectable among groups, while specific pairwise comparisons of effect sizes (using ΔV or other metrics) will allow discernment of the most pronounced differences in variance component. Regardless of the statistical approach used, we suggest it is important that researchers clearly outline the direction and, when possible, magnitude of the expected effects in their predictions.

Finally, our conceptual examples focus exclusively on the case of "well-behaved" data with normal error distributions. While these comparisons can be made with generalized extensions to LMMS (i.e. GLMMs), researchers must take extra precautions when calculating and comparing the within-individual variances (i.e. the residual variance). Indeed, in the case of non-Gaussian data, the residual variance depends on both the link function used and how the software deals with overdispersion (additive vs. multiplicative overdispersion). (Nakagawa and Schielzeth 2010)) provides a very useful and extensive guide explaining how the correct residual variation can be calculated.

CONCLUSIONS

Variance ratios are straightforward metrics to describe how various ecological and evolutionary processes occur. However, comparing these ratios across studies or group can be misleading if poor attention is given to the specific variance components making up those ratios. More importantly, a lack of difference in these ratios does not mean that variation is expressed equally among groups. Given these limitations, we advocate for techniques allowing the estimation of differences in each variance components rather than focusing solely on variance ratios. The statistical tools allowing comparison of trait variation have become increasingly sophisticated and now allow asking very precise questions. Specifically, we can now ask how trait variation is generated and how variation differs among groups. However, despite the availability of these tools, researchers interested in ecological and evolutionary variation must remain careful in their study designs. As our simulations show, scenarios involving differences in among-individual variance are particularly difficult to detect without substantial sample sizes. Finally, we hope the statistical approaches and tools for power analysis presented here will allow for appropriate comparisons of trait variation in ecological and evolutionary studies.

Acknowledgments

We thank the participants of the Statistical Quantification of Individual Differences (SQuID) Symposium at the 2016 ISBE Congress for helpful discussions. We also thank Russel Bonduriansky, Ben Bolker and four anonymous reviewers for helpful comments on a previous version of this manuscript. This study was funded by NSF IOS-1557951 (to NAD) and the Department of Biological Sciences at North Dakota State University.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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