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

ABSTRACT (256/350 words)

1. Many key questions in evolutionary ecology require the use of variance ratios such as heritability, repeatability, and individual resource specialization. These ratios allow 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 differs among populations and environments therefore often requires the comparison of these ratios across groups (i.e. populations, sexes, species).
2. 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 component among groups rather than in differences in variance ratios *per se*.
3. 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 Hierarchical Linear Modeling approaches (HLMs).
4. Frequentist and Bayesian frameworks have similar abilities to identify differences in variance components. However, variance differences at higher levels of organization (i.e. the among-unit variance) can be difficult to detect with low sample sizes.
5. 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

Running Head: Comparing variation within datasets

Keywords: Heritability, repeatability, individual niche specialization, animal personality, phenotypic variation, functional traits, mixed models, individual variation

INTRODUCTION

Our understanding of many evolutionary and ecological processes is underpinned by an estimation of variance ratios. For example, evolutionary change is dependent on the ratio of additive genetic variation (V_a) to total phenotypic variation (V_p), more commonly known as narrow-sense heritability ($\frac{V_a}{V_p}$ or h^2):

$$\Delta z = h^2 s \quad \text{(equation 1)}$$

where the change in a population's mean from one generation to the next (Δz) is based on the selection differential (s) and the trait's heritability (h^2) (breeder's equation, Lush 1937). Considerable effort has been directed toward estimating and comparing heritability estimates among taxa or among trait types (Mousseau and Roff 1987; Stirling et al. 2002; Dochtermann et al. 2019), with these comparisons sometimes used to argue that some traits are under greater selection than others (Mousseau and Roff 1987).

Variance ratios are similarly important across ecology. For example, individual resource specialization can be estimated as the proportion of variation in an individual's resource use relative to the species' total variation in resource use (Bolnick et al. 2002):

$$specialization = \frac{WIC}{TNW} \quad \text{(equation 2)}$$

where TNW is a species' total niche width (total resource variation) and WIC is "the average variance of resources found within individual's diets".

Interest in variance ratios spans a broad swath of evolutionary ecology (Table 1). This includes interest in repeatability and "animal personality" (Lessells and Boag 1987; Bell et al. 2009; Dingemanse and Dochtermann 2013; Dochtermann et al. 2015) and interest in community ecology regarding the distribution of functional trait variation expressed within versus among populations or species (Violle et al. 2012).

While the use of variance ratios can facilitate comparison among populations, inferences based on these ratios can be highly misleading (Houle 1992; Wilson 2018). If a variance ratio is compared between two groups, this comparison is only narrowly interpretable. Specifically, such a comparison is not informative regarding the biological basis of a difference or lack thereof. This is the case because variance ratios can differ when their numerators differ, their denominators differ, or because both differ. Indeed, variance ratios can be equal despite having different numerators and denominators values.

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 & Roff 1987
Behavioral Ecology	*Repeatability*	$\mathrm{R} = Vi / Vp$	The proportion of variation attributable to among-individual differences (Vi)	Lessels & Boag 1987
Ecology	*Individual Niche Specialization*	$\mathrm{S} = WIC / TNW$	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*	$\mathrm{T}_{\mathrm{IP/IC}} = V_{IP} / V_{IC}$	The proportion of variation attributable to within-population variance (V_{IP}) relative to the community variance (V_{IC})	Violle et al. 2012
		$\mathrm{T}_{\mathrm{IC/IR}} = V_{IC} / V_{IR}$	The proportion of variation attributable to community variance (V_{IC}) relative to the regional pool variance (V_{IR})	

Legend: Va: additive genetic variance in trait, Vi: among-individual variance in trait, Vp: total (i.e. phenotypic) variance in 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 illustrate that point further, let us consider the following scenario: researchers are studying the behaviors and dietary habits of two populations of the mythical Dahu (*Dahu desterus*; Figure 1A) at different elevations. These elusive creatures have shorter hind-legs on their left side, thus only allowing for clockwise movement (Chartois & Claudel 1945; Jacquat 1995). While measuring aggressive interactions, researchers find no differences in means between populations and similar behavioral repeatabilities (τ = 0.8; Figure 1B). 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. However, paying attention to the variance components reveals that individuals from the high-altitude population are much more distinct from one another in their aggressive tendencies while, at low-altitude, individuals show little departure from the population average (Figure 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 (Figure 1D) while the within-individual variation in niche preference is equal between populations. This means that it is the difference in diet preference among individuals that drives the difference between the two populations. With more diverse resources available at low elevation each individual can specialize along the total niche axis, yet the breadth of diet preference within-individuals is unchanged in both populations.

For both traits, exclusive reliance on ratios would have led to either inappropriate or incomplete inferences. Due to these problems with interpretations of variance ratios, what would be of greater use to researchers is to understand differences in the underlying variance components themselves.

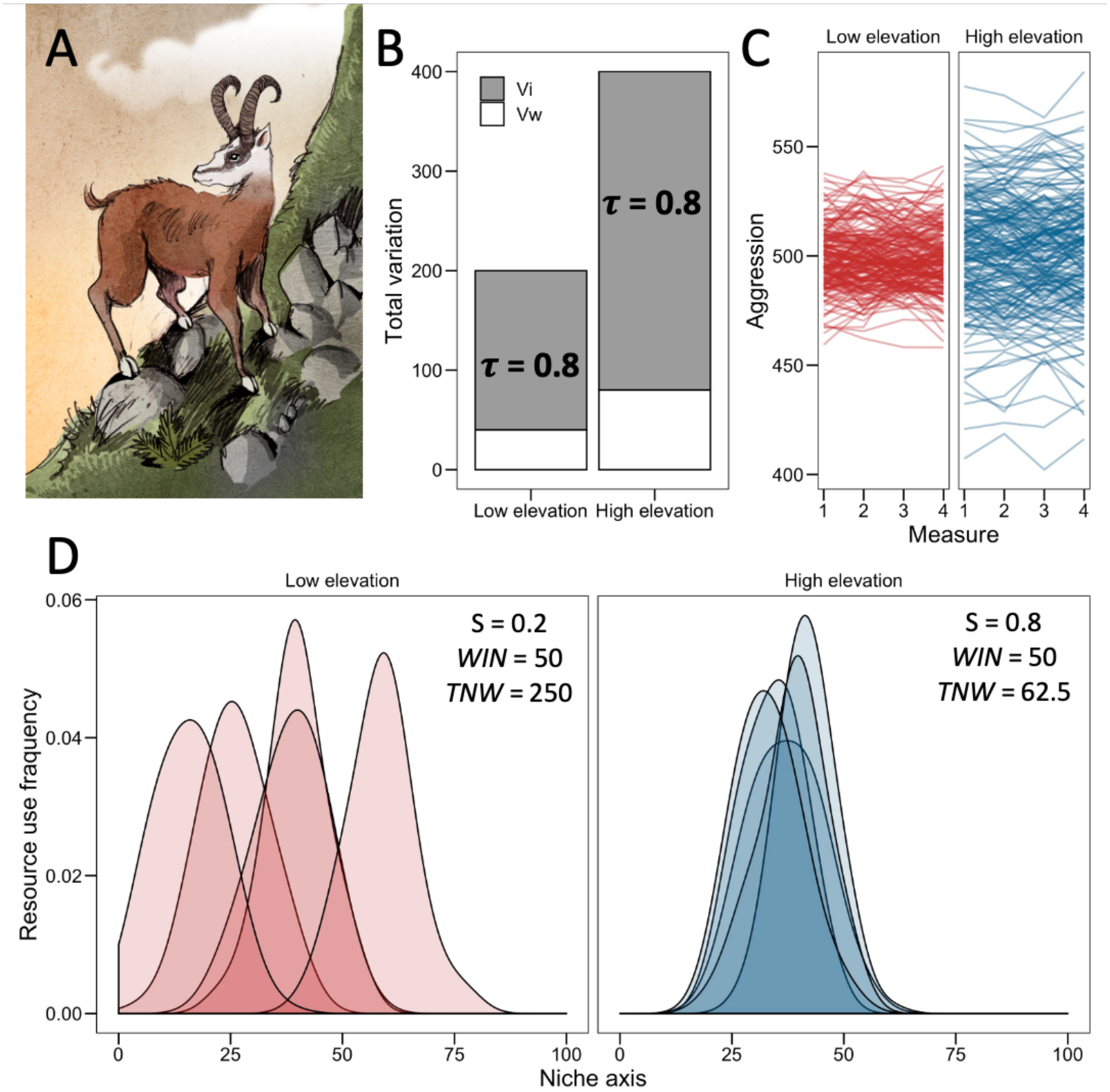

Figure 1. Reliance on variance ratios can lead to misleading inferences. (A) The elusive Dahu (*Dahu dexterus*) in its natural environment. (B) Two populations of Dahus living at different elevations do not differ in their repeatability of aggressive interactions (τ). (C) By plotting the individual aggression 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-altitude population. Figure code available here: https://osf.io/5aw42/

Illustration: Philippe Semeria (CC BY 3.0 license)

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 & 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 & Roff 2010 and White et al. 2019). 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.

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 experimental units (V_H and V_W respectively). We use the subscripts H and W to denote that the among-unit variance (V_H) represents the "higher-level" variance used for comparing differences between the two environments, while the within-unit variance (V_W) indicates differences in trait value occurring within environments during the course of the experiment. This is a broadly applicable scenario that can correspond to the comparison of the repeatability of a phenotype between environments, 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 (τ = $V_H/(V_H + V_W)$) is to estimate the variance components for each environment in separate statistical models. We can then test for differences in variance components and ratio by environments based on whether their confidence intervals overlap or not. While straightforward, this method suffers from several 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 α = 0.05 for the null hypothesis of no difference (Austin and Hux 2002; MacGregor-Fors and Payton 2013; Hector 2015). 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 Hierarchical Linear Model (HLM) where the among- and within-unit variance is estimated for each environment within the same statistical model. This statistical model can be described by the following equation:

$$y_{ij} = \beta_0 + \beta_1 Environment + unit_{0j} + e_{0ij} \qquad \text{(equation 3)}$$

$$unit_{0j} \sim MVN(0, \Omega_{unit}); \quad \Omega_{unit} = \begin{bmatrix} V_{unit0}\, E_1 & 0 \\ 0 & V_{unit0}\, E_2 \end{bmatrix}$$

$$e_{0ij} \sim MVN(0, \Omega_e); \quad \Omega_e = \begin{bmatrix} V_{e0}\, E_1 & 0 \\ 0 & V_{e0}\, E_2 \end{bmatrix}$$

where y_{ij} describes the phenotypic traits for the *i*th experimental unit and *j*th observation. $unit_{0j}$, is the deviation from an overall intercept, β_0, for the *j*th experimental unit. 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_{unit} and Ω_e, are the variance-covariance matrices at the among- and within-unit levels respectively.

The diagonal elements of these matrices represent the among- (H) and within-unit (W) variances by environment and the off-diagonal elements represent the cross-environment correlation (set to 0 if units 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). HLMs are now available for most statistical software and their generalized extensions can accommodate non-normal error distributions (Table 2).

Upon fitting HLMs, 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-unit variance differs by environment can be compared to a null model where the among and within-unit variance are kept constant across 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 2000; see Santostefano et al. 2016 for a recent example).

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 HLMs 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), 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 used ΔV on an additive scale because it allows the most straightforward interpretation and functions in cases where a variance component is zero or approaching zero.

Table 2. Packages and softwares allowing to test for differences in variance components using Hierarchical Linear Models (HLM) along with parameter estimation method (maximum likelihood (ML), restricted maximum likelihood (REML) or Bayesian framework) and inference method (Likelihood Ratio tests (LRT), AIC or credible interval overlap). 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	Within-unit variance by group	Distributions handled	Comments	Reference
ASREmL	Commercial	ML/REML	LRT, AIC	Yes	Yes	Gaussian		Gilmour et al. (2015)
SAS	Commercial	ML/REML	LRT, AIC	Yes	Yes	Gaussian, Poisson, Binomial …		SAS Institute Inc.
nlme	Free	ML/REML	LRT, AIC	Yes	Yes	Gaussian		Pinheirho and Bates (2000)
lme4	Free	ML/REML	LRT, AIC	Yes	No	Gaussian, Poisson, Binomial …		Bates et al. (2015)
R-INLA	Free	ML/REML	LRT, AIC	Yes	Yes	Gaussian		Lindgren, and Rue (2015)
glmmTMB	Free	ML/REML	LRT, AIC	Yes	Yes	Gaussian, Poisson, Binomial …		Brooks et al. 2017
hglm	Free	ML/REML	LRT, AIC	Yes	Yes	Gaussian, Poisson, Binomial …	Within-unit variance modelled as Gamma distribution	Rönnegård et al. (2010)
MCMCglmm	Free	Bayesian	DIC, overlap of credible intervals	Yes	Yes	Gaussian, Poisson, Binomial …		Hadfield (2010)
brms	Free	Bayesian	WAIC, LOO, overlap of credible intervals	Yes	Yes	Gaussian, Poisson, Binomial …	Within-unit variance modelled as log-normal distribution	Bürkner (2017)

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 HLM 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 HLMs to detect differences in variance components.

METHODS

Data simulations

To compare the performance of statistical procedures for the detection of differences in variance components and variance ratios, we performed a series of simulations based on the scenarios illustrated in Figure 2. In these scenarios a phenotypic attribute y is measured in two different environments (E1 and E2) and variation occurs among and within experimental units (V_H and V_W respectively). In scenarios A through C the variance ratio 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-unit variance (A), in the within-unit variance (B) or in both the among and within-unit 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 other variance components (E) (see Table S1 for exact values for all parameters).

Using the R statistical environment (R Core Team 2017), we generated 500 datasets for each of the following combinations:

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

Each dataset was simulated by sampling from a Gaussian distribution for the random (among-unit values) and the error (within-unit) 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 Supporting Information 1.

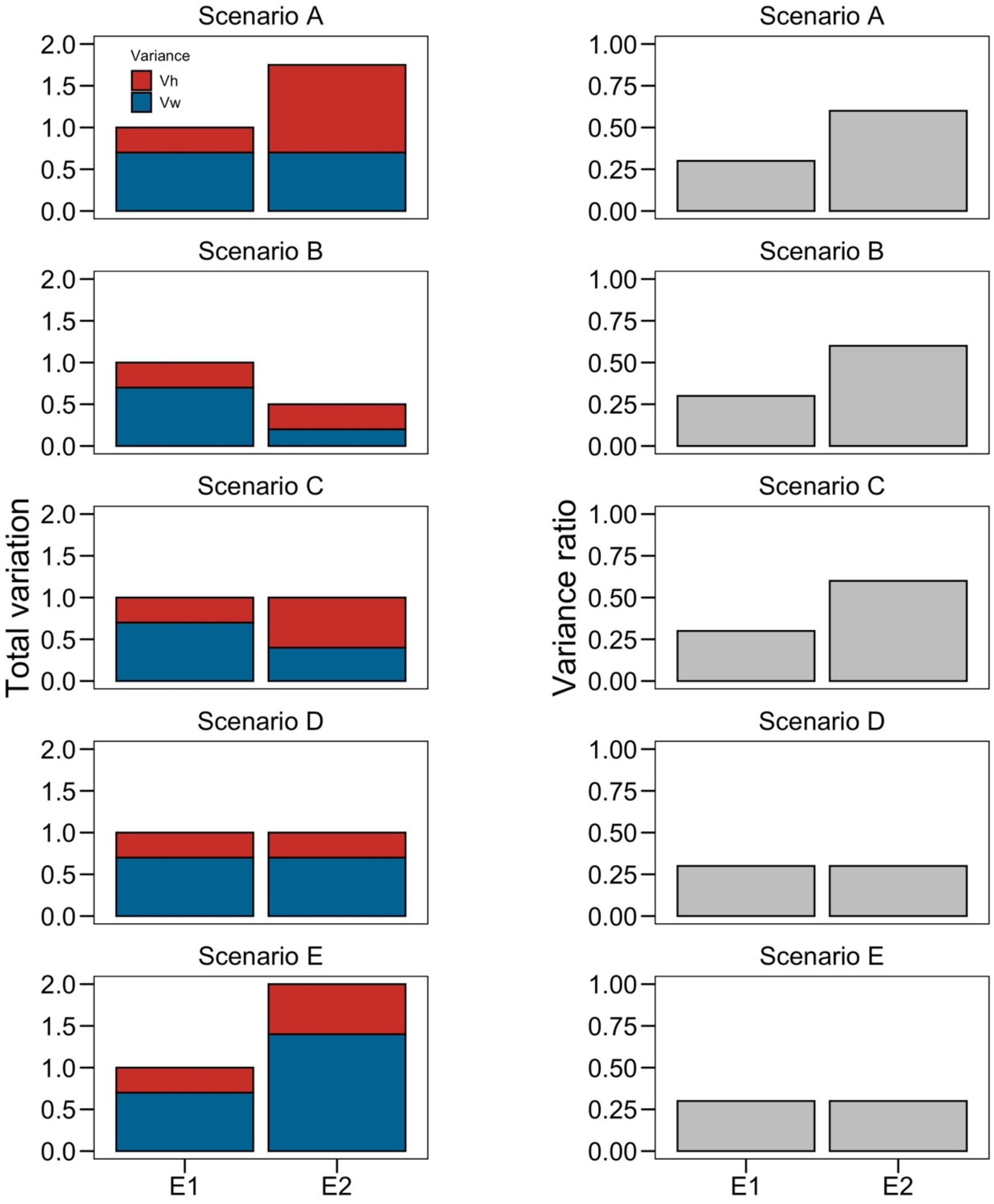

Figure 2. Scenarios used in simulations detailing how differences or lack of difference in variance ratios can arise from different patterns in the underlying variance components (Exact values can be found in Table S1). Scenarios A-C correspond to cases where the total variation differs between two environments (E1 and E2) due to differences in the higher group level variance (V_H, A), the lower level variance (V_W, B) or both (C). Scenarios D-E indicate cases where the ratios remains constant across environments, because all variance components are indentical (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):

$$y_{ij} = \beta_0 + unit_{0j} + e_{0ij} \qquad \text{(equation 4)}$$

$$unit_{0j} \sim \mathcal{N}(0, V_{unit});$$

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

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-unit variance. Datasets where these intervals did not overlap were considered as statistically different.

Frequentist HLM with AIC model comparison

Our second approach was to fit the HLM approach described above and test for the for the significance of the difference in among- and within-unit variance using likelihood ratio tests. Specifically, we compared the following models:

We specified four different mixed models corresponding to the four different possibilities by which variance components may differ (see also Royauté et al. 2019; Bucklaw and Dochtermann 2021):

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

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

For each dataset combination, we then compared each model's Aikaike's Information Criterion value (AIC). AIC allows to compare the 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 (e.g. Jenkins 2011; Shaw 1991; Tüzün et al. 2017). These models were specified using the *nlme* package for mixed models (Pinheiro et al. 2000) using Restricted Maximum Likelihood (REML).

Bayesian HLM 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 $\Delta V = V_{\mathrm{E2}} - V_{\mathrm{E1}}$) 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 SI1 and SI3 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 detecting differences), 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 detecting differences 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 detecting differences 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 (Figure 3). The highest probability to detect differences was observed for in cases where the variance ratio differs as a result of changes to the within-unit 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 (Figure 3). Importantly, no statistical method seemed to outperform all others across scenarios. Our results are consistent with previous simulations showing that the among-unit variance component is particularly difficult to estimate at small sample sizes (Dingemanse & Dochtermann 2013).

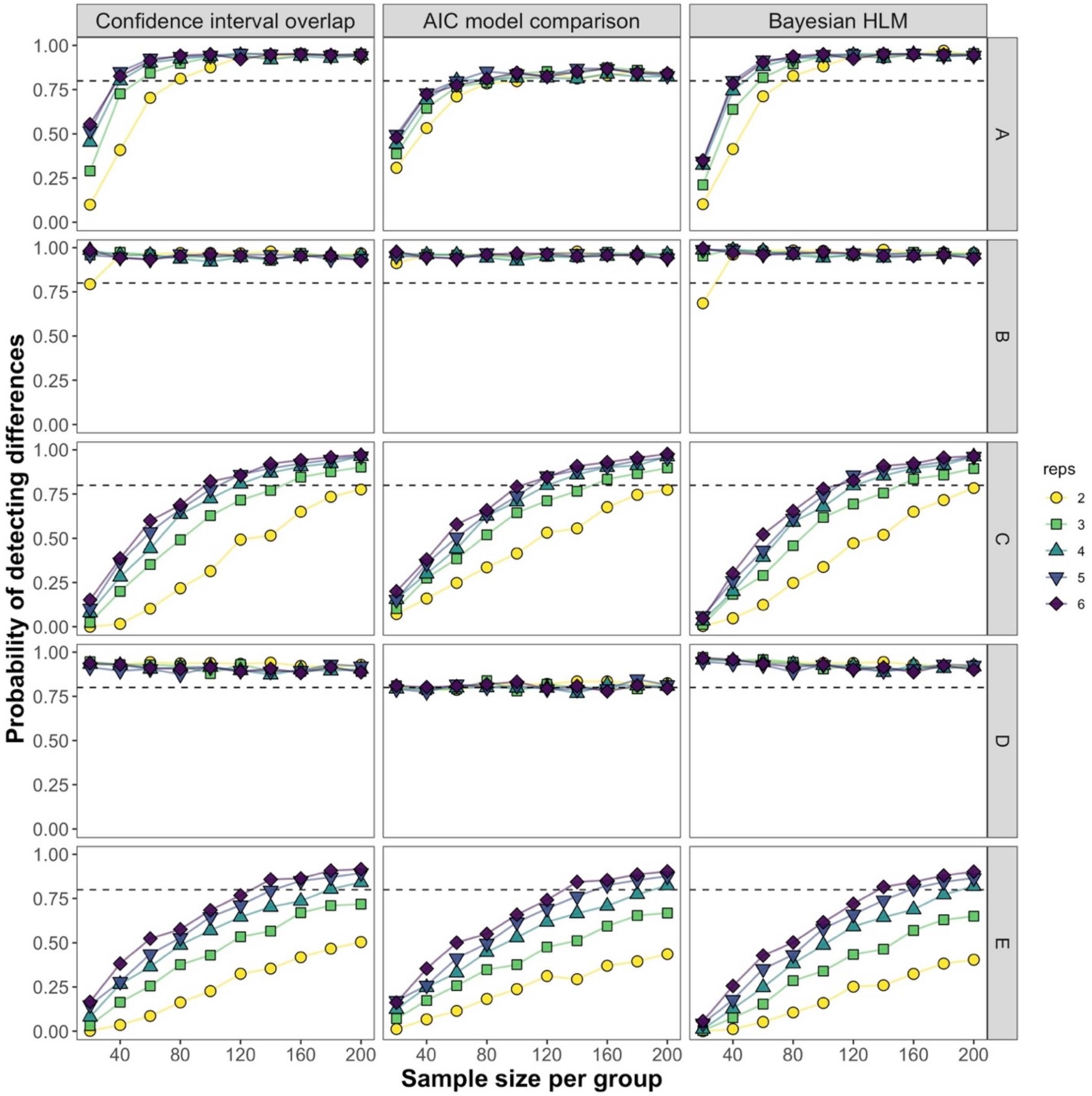

Figure 3. Effect of sampling design on the probability to detect differences in variance components by scenario type and statistical modeling approach. Each point represents the probability of detecting the correct differences in variance averaged over 500 simulated datasets. A represents a scenario where only the among-unit variance (V_H) varies between environments, B represents a case where the within-unit variance (V_W) varies between environments, and both among and within-unit 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 % treshold similar to recommendations for power analyses.

In scenarios B and D, the correct differences among variance components was identified > 80 % of the time, even at low sample sizes (Figure 3). In all other cases 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-unit variance (C) or when the variance ratio remains the same despite changes to both among- and within-unit variance (E) – datasets with only 2 repeated measures per unit never achieved a power above 0.8 even with sample sizes above 200 units per environment (i.e. a minimum of 800 total measurements, Figure 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 unit reached statistical power above 0.8 with sample sizes above 120 units per environments, 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 (Figure 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 is likely because when the generating model does not include differences in the within-unit 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 different 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 HLMs consistently underestimated the among-unit variance for scenarios in which the among-unit variance differed between environments (scenarios A, C, and E) resulting in a severe bias at small sample sizes (Figure S2). However, Bayesian HLMs also had higher precision and accuracy compared to maximum likelihood (Figure S3, S4). This means that Bayesian estimates tend to be consistently more conservative than maximum likelihood regarding the magnitude of the among-unit 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 ratios (Houle 1992; Hansen et al. 2011). Instead, many questions require the explicit comparison of variance components.

Our simulations show that regardless of the statistical methods used, comparing variance components across groups is a “data hungry” question. Scenarios where the among-unit variance differed between environments were particularly hard to detect at low sample sizes. Our objective was not to provide a full exploration of parameter space in order to define the proper sample sizes to detect differences of various magnitude for each variance component. Instead, we focused on a subset of scenarios that are likely to be common in ecology and evolution.

Given the issues discussed above, how should researchers interested in ecological and evolutionary variation design their studies and report their findings? Based on our simulations, the probability to detect differences in variance components will depend in large part on the ability to estimate the among-unit variance component (V_H). A simple rule for sampling can therefore be to estimate the sample size needed to detect the lowest among-unit 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. We also recommend that power calculations be conducted prior to the experiment whenever possible (see R code for *a priori* power analyses in SI2 and R Markdown tutorial in SI3).

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 SI4). 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 HLM framework as described by Nakagawa et al. (2015) (see also Carmona et al. 2016 and Fontana et al. 2018 for approaches relevant to functional trait diversity).

While we limited our conceptual example to comparisons between two environments, the HLM 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 HGLM can allow the comparison of indices of diet specialization within and among species. In both cases, model section 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 HLMS (i.e. HGLMs), extra care must be taken to appropriately estimate and compare the within-unit variance depending on the error distribution specified (Nakagawa & Schielzeth 2010).

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, as we have shown, 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-unit 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 two 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.

Data Availability Statement

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.

REFERENCES

Aguirre, J., E. Hine, K. McGuigan, and M. Blows. 2014. Comparing **G**: multivariate analysis of genetic variation in multiple populations. Heredity 112:21-29.

Arnold, S. J., and P. C. Phillips. 1999. Hierarchical comparison of genetic variance-covariance matrices. II. Coastal-inland divergence in the garter snake, *Thamnophis elegans*. Evolution 53:1516-1527.

Austin, P. C., and J. E. Hux. 2002. A Brief Note on Overlapping Confidence Intervals. Journal of Vascular Surgery 36:194–195.

Barr, D. R. 1969. Using confidence intervals to test hypotheses. Journal of Quality Technology 1:256–258.

Bates, D., M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, et al. 2015. Package 'lme4.'

Bell, A. M., S. J. Hankison, and K. L. Laskowski. 2009. The repeatability of behaviour: a meta-analysis. Animal behaviour 77:771–783.

Bolnick, D. I., R. Svanbäck, J. A. Fordyce, L. H. Yang, J. M. Davis, C. D. Hulsey, and M. L. Forister. 2002. The ecology of individuals: incidence and implications of individual specialization. The American Naturalist 161:1–28.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M. and B. M. Bolker 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal 9:378-400.

Bucklaew, A. and N.A. Dochtermann. 2021. The effects of exposure to predators on personality and plasticity. Ethology 127:158-165.

Bürkner, P.-C. 2017. brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software 80:1–28.

Burnham, K. P., and D. R. Anderson. 1998. Practical use of the information-theoretic approach. Pages 75–117 *in* Model Selection and Inference. Springer.

Carmona, C. P., F. de Bello, N. W. Mason, and J. Lepš. 2016. Traits without borders: integrating functional diversity across scales. Trends in ecology & evolution 31:382–394.

Chartois, J., & Claudel, C. 1945. Hunting the dahut: a french folk custom. The Journal of American Folklore 58:21-24.

Coblentz, K. E., A. E. Rosenblatt, and M. Novak. 2017. The application of Bayesian hierarchical models to quantify individual diet specialization. Ecology 98:1535–1547.

Dingemanse, N. J., and N. A. Dochtermann. 2013. Quantifying individual variation in behaviour: mixed-effect modelling approaches. Journal of Animal Ecology 82:39–54.

Dochtermann, N. A., and D. A. Roff. 2010. Applying a quantitative genetics framework to behavioural syndrome research. Philosophical Transactions of the Royal Society B-Biological Sciences 365:4013-4020.

Dochtermann, N. A., and R. Royauté. 2019. The mean matters: going beyond repeatability to interpret behavioural variation. Animal Behaviour 153:147–150.

Dochtermann, N. A., T. Schwab, M. Anderson Berdal, J. Dalos, and R. Royauté. 2019. The Heritability of Behavior: A Meta-analysis. Journal of Heredity.

Dochtermann, N. A., T. Schwab, and A. Sih. 2015. The contribution of additive genetic variation to personality variation: heritability of personality. Proceedings of the Royal Society B: Biological Sciences 282:20142201.

Fontana, S., M. K. Thomas, M. Moldoveanu, P. Spaak, and F. Pomati. 2018. Individual-level trait diversity predicts phytoplankton community properties better than species richness or evenness. The ISME journal 12:356.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, Sj. Welham, and R. Thompson. 2015. ASReml user guide release 4.1 structural specification. Hemel hempstead: VSN international ltd.

Hadfield, J. D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 33:1–22.

Hamilton, J. A., R. Royauté, J. W. Wright, P. Hodgskiss, and F. T. Ledig. 2017. Genetic conservation and management of the California endemic, Torrey pine (*Pinus torreyana* Parry): Implications of genetic rescue in a genetically depauperate species. Ecology and Evolution 7:7370–7381.

Hansen, T. F., C. Pélabon, and D. Houle. 2011. Heritability is not Evolvability. Evolutionary Biology 38:258.

Hector, A. 2015. The New Statistics with R: An Introduction for Biologists. 1st edition. Oxford ; New York, NY: Oxford University Press.

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

Jacquat, M. S. 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, S. H. 2011. Sex differences in repeatability of food-hoarding behaviour of kangaroo rats. Animal Behaviour 81:1155–1162.

Lessells, C. M., and P. T. Boag. 1987. Unrepeatable repeatabilities: a common mistake. The Auk 104:116–121.

Lindgren, F., and H. Rue. 2015. Bayesian spatial modelling with R-INLA. Journal of Statistical Software 63:1-25.

Lush, J. 1937. Animal Breeding Plans. Iowa State College Press, Ames, Iowa.

Martin, J. G., D. H. Nussey, A. J. Wilson, and D. Réale. 2011. Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. Methods in Ecology and Evolution 2:362–374.

MacGregor-Fors, I., and M. E. Payton. 2013. Contrasting Diversity Values: Statistical Inferences Based on Overlapping Confidence Intervals. PloS One 8, no. 2. http://dx.plos.org/10.1371/journal.pone.0056794.

Mousseau, T. A., and D. A. Roff. 1987. Natural selection and the heritability of fitness components. Heredity 59:181.

Nakagawa, S., R. Poulin, K. Mengersen, K. Reinhold, L. Engqvist, M. Lagisz, and A. M. Senior. 2015. Meta-analysis of variation: ecological and evolutionary applications and beyond. Methods in Ecology and Evolution 6:143–152.

Nakagawa, S., and H. Schielzeth. 2010. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. Biological Reviews 85:935–956.

Nakagawa, S., and H. Schielzeth. 2012. The mean strikes back: mean–variance relationships and heteroscedasticity. Trends in Ecology & Evolution 27:474–475.

Pinheiro, J., and D. Bates. 2000. Mixed-Effects Models in S and S-PLUS. Springer Science & Business Media.

Roff, D. 2002. Comparing **G** matrices: A MANOVA approach. Evolution 56:1286-1291.

Roff, D. A., J. M. Prokkola, I. Krams, and M. J. Rantala. 2012. There is more than one way to skin a **G** matrix. Journal of Evolutionary Biology 25:1113-1126.

Rönnegård, L., X. Shen, and M. Alam. 2010. hglm: A package for fitting hierarchical generalized linear models. The R Journal 2:20–28.

Royauté, R., C. M. Buddle, and C. Vincent. 2015. Under the influence: sublethal exposure to an insecticide affects personality expression in a jumping spider. Functional Ecology 29:962–970.

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

Royauté, R., C. Garrison, J. Dalos, M. A. Berdal, and N. A. Dochtermann. 2019. Current energy state interacts with the developmental environment to influence behavioural plasticity. Animal Behaviour 148:39–51.

Santostefano, F., A. J. Wilson, Y. G. Araya-Ajoy, and N. J. Dingemanse. 2016. Interacting with the enemy: indirect effects of personality on conspecific aggression in crickets. Behavioral Ecology 27:1235–1246.

Shaw, R. G. 1991. The comparison of quantitative genetic-parameters between populations. Evolution 45:143-151

Stirling, D. G., D. Réale, and D. A. Roff. 2002. Selection, structure and the heritability of behaviour. Journal of Evolutionary Biology 15:277–289.

Tüzün, N., S. Müller, K. Koch, and R. Stoks. 2017. Pesticide-induced changes in personality depend on the urbanization level. Animal behaviour 134:45–55.

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 in Ecology and Evolution 3:268–280.

Violle, C., B. J. Enquist, B. J. McGill, L. I. N. Jiang, C. H. Albert, C. Hulshof, V. Jung, et al. 2012. The return of the variance: intraspecific variability in community ecology. Trends in ecology & evolution 27:244–252.

White, S. J., Pascall, D. J., and A. J. Wilson. 2019. Towards a comparative approach to the structure of animal personality variation. Behavioral Ecology.

Wilson, A. J., D. Réale, M. N. Clements, M. M. Morrissey, E. Postma, C. A. Walling, L. E. B. Kruuk, et al. 2010. An ecologist's guide to the animal model. Journal of Animal Ecology 79:13–26.

Wilson, A. J. 2018. How should we interpret estimates of individual repeatability? Evolution Letters *2:* 4-8.

Supporting Information

SI 1: Zip folder containing the raw data from simulations along with R code for data analysis and figures (https://osf.io/5aw42/).

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

SI 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.

Figure 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.

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

Figure S3. Effect of sampling design on estimate precision (width of the interquartile interval) by scenario type and statistical modeling approach.

Figure S4. Effect of sampling design on model accuracy (estimated as the root mean square of error, RMSE) by scenario type and statistical modeling approach.