Behavioral Ecology and Sociobiology (2024) 78:18
https://doi.org/10.1007/s00265-024-03435-w

METHODS PAPER

The effect of repeated measurements and within-individual variance on the estimation of heritability: a simulation study

Mónika Jablonszky[1] · László Zsolt Garamszegi[1]

Received: 2 September 2023 / Revised: 12 January 2024 / Accepted: 22 January 2024 / Published online: 2 February 2024

Abstract

The estimation of heritability is a common practice in the field of ecology and evolution. Heritability of the traits is often estimated using one single measurement per individual, although many traits (especially behavioural and physiological traits) are characterized by large within-individual variance, and ideally a large number of within individual measurements can be obtained. Importantly, the effect of the within-individual variance and the rate at which this variance is sampled on the estimation of heritability has not been thoroughly tested. We fill this gap of knowledge with a simulation study, and assess the effect of within- and between-individual sample size, and the true value of the variance components on the estimation of heritability. In line with previous studies we found that the accuracy and precision of heritability estimation increased with sample size and accuracy with higher values of additive genetic variance. When the sample size was above 500 accuracy and power of heritability estimates increased in the models including repeated measurements, especially when within-individual variance was high. We thus suggest to use a sample of more than 100 individuals and to include more than two repeated measurements per individual in the models to improve estimation when investigating heritability of labile traits.

Significance statement

Heritability reflects the part of the trait's phenotypic variation underlined by genetic variation. Despite the difficulties of heritability calculation (high number of individuals is needed with known relatedness), it is a widely used measure in evolutionary studies. However, not every factor potentially affecting the quality of heritability estimation is well understood. We thus investigated with a comprehensive simulation study how the number of repeated measurements per individuals and the amount of within-individual variation influence the goodness of heritability estimation. We found that although the previously described effect of the number of studied individuals was the most important, including repeated measurements also improved the reliability of the heritability estimates, especially when within-individual variation was high. Our results thus highlight the importance of including repeated measurements when investigating the heritability of highly plastic traits, such as behavioural or physiological traits.

Keywords Animal model · Narrow-sense heritability · Quantitative genetics · Simulation · Variance components

Introduction

Determining how much genetic variance is present in phenotypic traits is a crucial step in understanding their adaptive evolution (Fisher 1930; Mousseau and Roff 1987). There are multiple estimates used to assess the evolutionary potential of traits in a population. The most frequently calculated measure is heritability (Mousseau and Roff 1987; Postma 2014), although evolvability may be more adequately measured with the mean-standardized additive genetic coefficient of variation (Houle 1992). There are many difficulties in estimating heritability correctly and precisely. This process

Communicated by J. Lindström

✉ Mónika Jablonszky
jablonszky.monika@gmail.com

[1] Evolutionary Ecology Research Group, Institute of Ecology and Botany, HUN-REN Centre for Ecological Research, Vácrátót, Hungary

requires large sample size and reliable data on the relationships between the individuals (Quinn et al. 2006; Morrissey et al. 2007; de Villemereuil et al. 2013). These requirements are especially difficult to be fulfilled in natural populations, in spite of that results from the wild are essential when studying evolution (Kruuk and Hadfield 2007; Postma 2014).

Animal models are widely used for estimating heritability of different traits, including labile traits, such as behaviour (Stirling et al. 2002; Kruuk 2004; Postma 2014). These models decompose additive genetic variances and environmental variances based on pedigree or other relatedness data (e.g. genetic similarity), and they are very flexible in controlling for confounding effects (e.g. dominance, common environment, maternal effects) (Wilson et al. 2010). Furthermore, if repeated measurements from the same individuals are included in the animal model, it can also discern permanent environmental variance (fixed differences between the individuals due to environmental and/or non-additive genetic effects) apart from additive genetic and residual variance (Kruuk 2004; Wilson et al. 2010). In addition to the additive genetic and residual variance, determining the amount of permanent environmental variance is also essential in predicting the evolutionary response of a trait.

Simulations are very important source of information for planning studies and assessing the reliability of studies investigating heritability. Simulations revealed that the sample size (de Villemereuil et al. 2013; Krag et al. 2013), the amount of the true heritability (Charmantier and Réale 2005; de Villemereuil et al. 2013; Krag et al. 2013), the type (genetic or social) (Bourret and Garant 2017) and the quality of the relatedness data (Israel and Weller 2000; Charmantier and Réale 2005; Kruuk and Hadfield 2007; Morrissey et al. 2007; de Villemereuil et al. 2013; Bourret and Garant 2017), structure of the simulated population (Clément et al. 2001; Kominakis 2008), data missing non at random (Steinsland et al. 2014) and also the analytical method (Kruuk and Hadfield 2007; de Villemereuil et al. 2013) can influence heritability estimates.

However, in spite of the huge amount of simulation research on the estimation of heritability (Clément et al. 2001; Morrissey et al. 2007; Bourret and Garant 2017), some aspects of this issue remained less explored. The calculation of heritability may be complicated by the remarkable within-individual variance that is characteristic of many behavioural, physiological and life history traits (Bell et al. 2009; Schoenemann and Bonier 2018; Taff et al. 2018). Within-individual variance has important biological significance, as it determines how well the individual can adapt to the changing environmental conditions, which is especially important during the recent climate change (Charmantier and Gienapp 2014). Moreover, within-individual variance has essential influence on the evolution of the traits, as it can promote or hinder adaptation (Piersma and Drent 2003; Snell-Rood 2013). However, there are simulation studies, showing how low repeatability (large within individual variance) influences the estimation of statistical parameters with evolutionary relevance, as it can induce bias in e.g. among-individual and residual variance (Schielzeth et al. 2020). Importantly, the large within-individual variance of labile traits relative to among-individual variance leads to small repeatability, which can be the upper limit of heritability (but see: Dohm 2002); thus, heritability is also expected to be small (see also: Mousseau and Roff 1987; Weigensberg and Roff 1996; Stirling et al. 2002). It was found repeatedly that it is more difficult to precisely and accurately estimate lower heritability (Klein 1974; Krag et al. 2013). However, it is crucial to estimate these small heritabilities precisely, as for example, in the song of the collared flycatcher (*Ficedula albicollis*) we have seen that revealing small but non-zero heritability can have strong theoretical implications, as it still can be the base of evolution (Jablonszky et al. 2022). Importantly, in spite of the well-known effect of the amount of heritability (Klein 1974; Charmantier and Réale 2005; Raffa and Thompson 2016), the effect of the amount of within-individual variance on the heritability estimates has not been thoroughly tested. Because heritability is estimated based on variance components, we can assume similar responses as the above mentioned effects during the estimation of among- and within-individual variances (Schielzeth et al. 2020). In previous studies, the among-individual variance was biased upwards and residual variance was usually biased downwards with low repeatability (Schielzeth et al. 2020). Thus, we predict less accurate (specifically upwardly biased) and less precise heritability estimates when the within-individual variance is large, especially at low sample sizes.

Another factor potentially influencing heritability estimation that received less attention is the number of repeated measurements included in the models. Collecting repeated measurements is common practice during the investigation of labile traits. Including the mean of these repeated measurements into animal models is not appropriate (Wilson et al. 2010; Ge et al. 2017; Risk and Zhu 2018; see also Garamszegi 2016 for other types of models) as the within-individual variance is removed from the variance components (i.e. the uncertainty around the mean estimate is not accounted for) resulting in upwardly biased estimates (Åkesson et al. 2008; Hadfield et al. 2010; Silva et al. 2017). Thus, all repeated measurements should be included in the models (Wilson et al. 2010). Although information on additive genetic variance comes from the data on the relatedness among the individuals (thus, the reliability of the estimation depends primarily on the number of individuals), the estimation of other variance components taking part in heritability calculation, such as within-individual variance, are sensitive to the number of repeats used (Royauté and Dochtermann 2021). Thus, it would be worthwhile to investigate whether collecting more measurements from the individuals improve heritability estimation. More repeats mean higher overall sample size, but also cover a wider range of possible trait values (until a certain level of sampling), resulting in better estimation of all variance components (Westneat et al. 2020), so we can expect more precise and accurate heritability estimates with higher number of repeated measurements. Although, if the number of repeated measurements is too low, additive genetic and permanent environmental effects cannot be separated reliably (Bourret and Garant 2017). However, the effect of the number of repeated measurements, especially in the case of labile traits, has received less attention in simulation studies, although we know that repeated measurements can increase power in linkage analyses (Zhang and Zhong 2006; Liang et al. 2009). We are aware of only one study, that showed reduced uncertainty around the estimates with increasing number of repeated measurements (Adams 2014). Additionally, the effect of within-individual variance and the number of repeated measurements could interact. Previously, it was shown that estimation problems arising from low repeatability can be eliminated if appropriate number of repeated measurements of the same individuals is included in the models (Martin et al. 2011; Dingemanse and Dochtermann 2013; Westneat et al. 2020). Similarly, probably more repeated measurements are necessary for the correct separation of permanent environmental and residual effects when within-individual variance is large (Martin et al. 2011).

In this simulation study, to fill the abovementioned gaps in our knowledge, our aim was to investigate the effect of increasing within-individual variance at different combinations of within- and between-individual sample sizes in the animal models. Specifically, we simulated datasets to investigate the effect of different amount of variances (we varied the value of additive genetic variance, permanent environmental and within-individual variance from small to large, between 0.1–0.5, 0–0.8 and 0.1–0.8, respectively), as well as within- and between-individual sample sizes (1–10 repeats from 100 to 1000 individuals) on the estimation of heritability. We compared the error, accuracy, precision and power of these scenarios (see details in the "Methods" section).

Methods

Data simulation

We simulated datasets with all combinations of number of individuals (Ni = 100, 500, 1000) and number of measurements (Nr = 1, 2, 5, 10). The simulated value for the additive genetic variance (Va) and the residual variance (Ve) were 0.1, 0.3 or 0.5, and we also simulated a scenario, when Va was 0.1 and Ve was 0.8 to cover the feasible range of the values (resulting in 10 different scenarios). Thus, we had 120 different scenarios based on the combination of different parameter settings. Residual variance usually represents the combined effect of within-individual variance, measurement error, and unaccounted environmental variance, but as we included no measurement error and environmental variance into our simulated data, we will regard this component (Ve) as within-individual variation in the followings. Permanent environmental effect (Vpe) was simulated in a way that the sum of variance components became 1 and thus its value was between 0 and 0.8. In the models with only one measurement per individual Vpe and Ve is summed and represent the residual variance together (later we refer to this term as Vr).

We simulated 100 datasets for each scenario. Running more rounds was not feasible because of the large number of scenarios and the high computational demands of the Bayesian models. As a first step, we built a pedigree in each simulation with the 'generatePedigree' function from the 'geneticsPed 1.56' package (Gorjanc et al. 2021). For all scenarios, the pedigree was simulated for the appropriate number of individuals with 5 generations (thus the number of individuals per generations were Ni/5), and with Ni/25 dams and sires per generation. For simplicity the simulated population was assumed to be closed, with complete random mating and non-overlapping generations. To check the effect of pedigree structure on our results, we run additionally some scenarios with different parameters for pedigree construction (5 generations, but Ni/2 dams and sires, 3 generations, lower number of sires than dams), but these settings did not influenced our results qualitatively (see Tables S2–S4). Additive genetic component was simulated with the 'rbv' function from the 'MCMCglmm 2.32' package (Hadfield 2010), using the appropriate value of additive genetic variance for the scenario. Permanent environmental effect was simulated for all individuals and the within-individual term was calculated for all measurements with the corresponding consideration for these variance components. All of these effects were assumed to be normally distributed. The phenotypic value for each individual was the sum of the population mean (which was arbitrarily assigned to the value of 1), additive genetic, permanent environmental and within-individual components:

$$y_{ij} = \mu + a_i + p_i + e_{ij} \tag{1}$$

$$a_i \sim N(0, Va)$$

$$p_i \sim N(0, Vpe)$$

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

where y_{ij} is the phenotype of the ith individual at the jth repeat, μ is the population mean, a_i is the additive genetic effect, p_i is the permanent environmental effect and e_{ij} is the within-individual effect of the *i*th individual at the *j*th repeat.

Analysis of the simulated datasets

On the generated data we run animal models with the 'MCMCglmm' function from the 'MCMCglmm 2.32' package (Hadfield 2010). The models for the scenarios with only one measurement per individual contained only one random factor of individual identity connected to the pedigree:

$$y_i = mu + a_i + r_i \tag{2}$$

$$a_{\mathrm{i}} \sim N(0, Var(\mathrm{a}))$$

$$r_{\mathrm{i}} \sim N(0, Var(\mathrm{r}))$$

where y_{ij} is the estimate for the phenotype of the *i*th individual, mu is the estimate for the population mean, a_i is the estimate of the additive genetic effect and r_i is the residual effect of the *i*th individual. We use here r_i as this term include both permanent environmental and within-individual effects.

For the datasets with repeated measurements, we built models with two random factors for individual identity to separate additive genetic and permanent environmental effects:

$$y_{ij} = mu + a_i + p_i + e_{ij} \tag{3}$$

$$a_{\mathrm{i}} \sim N(0, Var(\mathrm{a}))$$

$$p_{\mathrm{i}} \sim N(0, Var(\mathrm{p}))$$

$$e_{\mathrm{ij}} \sim N(0, Var(\mathrm{e}))$$

where y_{ij} is the estimate for the phenotype of the *i*th individual at the *j*th repeat, *mu* is the estimate for the population mean, a_i is the estimate of the additive genetic effect, p_i is estimate of the permanent environmental effect and e_{ij} is the estimate for the within-individual effect.

Priors with inverse-Gamma distribution were used for all models. However, we checked the effect of other priors (e.g. parameter expanded prior) for some scenarios and results remained qualitatively unchanged. The models were run for 110,000 iterations with 10,000 sample discarded at the beginning and a thinning intervals of 100. Before running all simulation, the trace and distribution of all variables and the autocorrelation between iterations were checked visually for some selected scenarios.

From all models, the median of the estimate of heritability (the median of additive genetic variance divided by the sum of all variance components, hereafter h^2) and the variance components with their 95% credible intervals (CI) based on the whole posterior distributions were extracted with 'HPDinterval()'. We did not use posterior mode as it was proved to be prone to bias (Pick et al. 2022). To assess whether our heritability estimates can be differentiated from that of a scenario with zero heritability, we also calculated for the h^2 estimates the percentage of the values of a posterior distribution from a null model (run on a null dataset with Va = 0) that were greater than the actual estimates (Pick et al. 2022). We simulated one null dataset for all scenarios in a similar way as described above but with Va = 0 and Ve = $\mathrm{Va}_{\mathrm{actual}}$ + Ve ($\mathrm{Va}_{\mathrm{actual}}$ is the Va of the focal scenario) to ensure the same overall variance (Pick et al. 2022). The null model was built for this null dataset in the same way as for the original dataset.

Performance metrics

Measures of estimation error, accuracy, precision and statistical power were calculated for all scenarios for the heritability estimates and the first three measures also for the variance components (these latter results can be seen in the Supplementary material Figs. S1–S3). Specifically, we measured measurement error as the root mean square error (RMSE), and accuracy as absolute relative bias (we used the specific terms hereafter). RMSE (a measure of estimation error, often termed as accuracy, but reflecting also precision) is the square root of the average squared difference of the generating value of the actual parameter (p) and the estimated parameter ($\hat{p}$ and n is the number of simulated datasets) (as used in de Villemereuil et al. 2013; Schielzeth et al. 2020):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{p}_i - p)^2}{n}} \tag{4}$$

Thus, we obtained one value for each scenario reflecting the average difference of the estimates from the original simulated values. High values indicate high estimation error, and values close to zero indicate good estimation.

Accuracy was assessed as the absolute relative bias (Pick et al. 2022):

$$relative\ bias = \frac{1}{n} \sum_{i=1}^{n} \frac{\left|(\hat{p}_i - p)^2\right|}{p} \tag{5}$$

Thus, accuracy also resulted in one averaged value per scenario.

Precision was calculated as the inverse of the standard deviation of the heritability estimates of each run of the scenario (as used in Pick et al. 2022). The distribution of the point estimates reflects the expected distribution of the heritability estimates of 100 replicated studies.

Statistical power was assessed by comparing our estimates to the estimates of a null model, see above. Specifically, we calculated the ratio of the h^2 values of a posterior distribution from a null model that were greater than the original estimates. Then, statistically power was equal to the ratio of the simulations when the above mentioned ratio was lower than 0.05 (these estimates will be referred as significant). Note that all performance estimates resulted in one value per each scenario.

All statistical analyses were performed in the R 3.6.1 statistical environment (R Core Team 2019).

Results

The RMSE values for the h^2 estimates were the highest (indicating bad performance) when the number of individuals was 100 (Fig. 1, first row). The RMSE values became much smaller on average by 25% when 10 measurements were included instead of one, but at $N_i = 100$ only when Va = 0.5, Vpe was 0.2 or 0 and Ve was 0.3 or 0.5 (30 and 80% decrease, respectively, Fig. 1a). Even in these cases, using 10 measurements did not have an advantage over using 5 measurements. Apart from these scenarios RMSE was influenced by the magnitude of Va: scenarios with higher true value of Va had higher RMSE. There was even a 2.65-fold increase in RMSE between Va = 0.1 and Va = 0.5 scenarios when $N_i = 100$ (Fig. 1a).

Precision was also low at $N_i = 100$ but showed different patterns when the between-individual sample size was higher (Fig. 1, second row). At $N_i = 500$, precision was the highest when Va = 0.1 and precision dropped sharply by 70 and 80% respectively for the scenarios where Ve was 0.3 or 0.8 if even one repeated measurement were included (Fig. 1e). However, in the scenario of $N_i = 500$, Va = Ve = 0.5 and Vpe = 0 precision of h^2 estimates increased 4.59-fold when 10 measurements was included instead of one. This scenario displayed the same behaviour also when N_i was 1000, along with the scenario of Va = 0.1 and Ve = 0.8 (Fig. 1f). In these cases 10 measurements resulted in better precision than 5.

Regarding relative bias, using at least 2 measurements caused significant improvement at $N_i = 100$, Va = 0.5, Vpe = 0.2 or 0 and Ve = 0.3 or 0.5 relative to the models with only one measurement (40 and 75% decrease in relative bias, respectively, Fig. 1g). At $N_i = 500$, relative bias decreased when 10 measurements was included instead of one on average by 25% and showed a marked decrease of 60% in the scenario where Va = 0.1 and Ve = 0.8 (Fig. 1h). At $N_i = 1000$, more scenarios with Va = 0.1 showed decreasing tendency of relative bias with the number of measurements (Fig. 1i). Some scenarios among all sample sizes showed very slightly increased bias when only two measurements were included in the models compared to the one measurement model.

The statistical power to detect significant h^2 estimates increased on average by 40% when 10 measurements was included instead of one (Fig. 1, fourth row). This increase depended also on the magnitude of the Ve component: it was higher when Ve increased (5% increase when Ve = 0.1 and 800% increase when Ve = 0.8 across all sample sizes and scenarios for the other variance components). The improvement of power relative to models with one measurement was as high as 161% for models with 2 measurements at $N_i = 100$ (Fig. 1j), but 5 measurements provided additional advantage when N_i was higher (but only when Va = 0.1 (an increase of 47%), because the other scenarios have very high power (80–100%) with these higher sample sizes, Fig. 1k, l).

Additionally, the exact value for all performance estimates for all scenarios (Table S1) and the mean, the standard deviation and the average 95% CI width of the estimates for h^2 and the variance components (Table S5) can be seen in the Supplementary material. In Table S5, we can see that heritability is usually underestimated. However, it is overestimated in most of the Va = 0.5, Vpe = 0.2, Ve = 0.3 scenarios with repeated measurements (with the exception of the Ni = 1000 and Nr = 10 scenario), and half of the Va = 0.3, Vpe = 0.2 and Ve = 0.5 scenarios with repeated measurements. In the one measurement models, if biased, Va was under- and Vr was overestimated. In the models with repeated measurements the bias came from the bad separation of Va and Vpe (usually underestimation of Va and overestimation of Vpe, except the above-mentioned exception where the pattern was reversed) as Ve was estimated relatively well in these models.

Overall, the scenario of Va = Ve = 0.5 and Vpe = 0 has the less bias under all sample size scenarios and the highest precision (if number of measurements was at least five). The other scenarios with Va = 0.5 and scenarios with Va = 0.3 at Ni = 500 or 1000 have also low relative bias, but did not show higher precision than the rest of the scenarios.

Discussion

Our simulation results highlight the need for considering the collection of repeated measurements when investigating heritability. In most of the scenarios using at least two measurements offered some advantage over using only one measurement in terms of accuracy and/or precision. For instance, in the scenario of Ni = 100, Va = Ve = 0.5, relative bias decreased by 75% and precision showed a 2.43-fold increase when having at least two measurements. Within-individual variance also should be taken into account when planning studies on heritability, as the magnitude of this variance component influenced the effect of the repeated

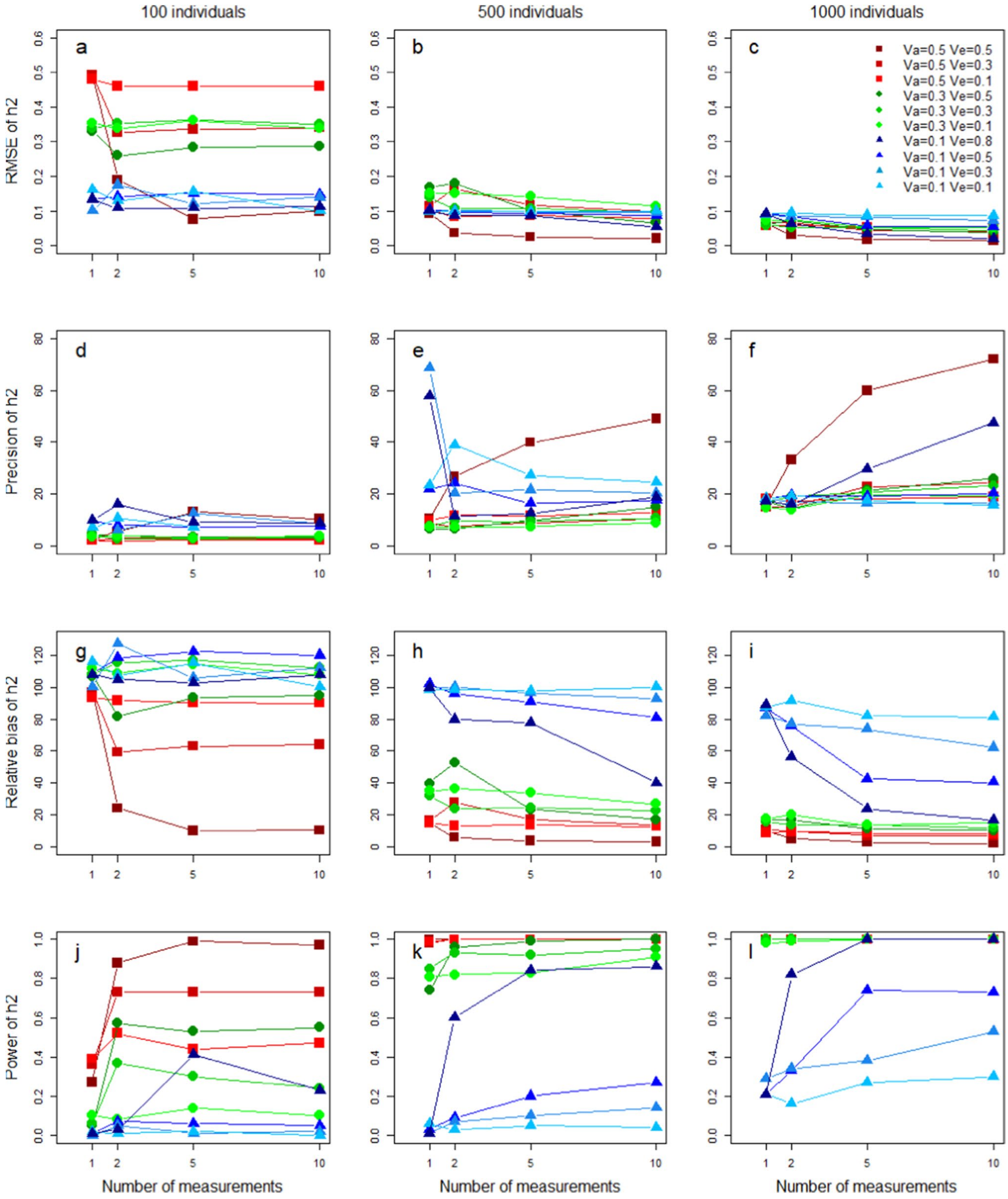

Fig. 1 Root mean square error (RMSE), relative bias, precision and power for the heritability (h^2) estimates are displayed for all scenarios, separately for the models with 100, 500 and 1000 between-individual sample size. The corresponding additive genetic variance (Va) values used in the simulations are depicted by colours and point types and within-individual variance (Ve) by shades of the respective Va value as can be seen in the legend. Vpe values were simulated in a way that the sum of all variance components became one

measurements on RMSE, relative bias, precision and statistical power of the heritability estimates. Models with 500 or 1000 individuals usually yielded estimate with low RMSE and relative bias and high precision, apart from scenarios with low heritability ($h^2 = 0.1$), where bias was significantly higher and power was lower. Biased estimates were usually underestimated. Although using 100 individuals seems to be insufficient to estimate heritability reliably, taking repeated measurements when the between-individual sample size is higher can increase accuracy and power (and sometimes also precision), especially in highly labile traits (i.e. high Ve).

The heritability estimates of the models including only one measurement per individual were influenced by the between-individual sample size and the magnitude of the true heritability. Generally, the models with one measurement yielded precise heritability estimates with low RMSE and relative bias at a sample size of 500 or 1000 individuals (aside from high relative bias for some scenarios with Va = 0.1, see below). These results were expected based on the sample size recommendations of at least 200, but possibly 300–1000 individuals of previous studies (Quinn et al. 2006; de Villemereuil et al. 2013; Krag et al. 2013). The accuracy and precision of heritability estimation also depended on the true heritability value, in a similar way as was found previously. In a comprehensive simulation study using 200 or 1000 individuals, the true value of heritability (0.1, 0.3, 0.5) also influenced the RMSE of the heritability estimates as estimates had less estimation error (i.e. lower RMSE) at 0.1 heritability (de Villemereuil et al. 2013). Another simulation study found higher bias for 0.1 than for 0.4 true heritability values when relying on 20–100 broods as sample size (Charmantier and Réale 2005), and these results also generally agree with our findings related to relative bias. Note that the trend in RMSE and relative bias according to the true heritability value was opposite both in previous papers and in our study. This emphasizes the need to investigate multiple performance metrics in simulation studies. We also investigated precision, and we found that heritability estimates were generally more precise when their generating value was low (thus, the previously mentioned RMSE values may reflect the higher precision of the estimates). However, higher precision for lower heritability may be only the consequence of that variance components are bound to be positive (de Villemereuil et al. 2013; Krag et al. 2013). Nevertheless, Krag et al. (2013) demonstrated based on simulations that for reliable estimates of heritability over 0.15 sample sizes larger than 400 individuals are needed. However, in our study, we found that the estimation of heritability of 0.1 can have still high (usually downward) relative bias and low statistical power with 500 or 1000 between-individual sample sizes. This fact is important to consider, as for example regarding behavioural traits, heritability estimates are often low, but sample size is usually below 1000 (heritability estimates were between 0.05 ± standard error: 0.02 and 0.21 ± 0.07, and number of individuals between 81 and 455 in the following papers: Blumstein et al. 2010; Santostefano et al. 2017; Jablonszky et al. 2022). Regarding life history traits, also low heritability estimates (0 ± 0.01 or 0.11 ± 0.003) were reported when investigating more than 1000 individuals (Brommer et al. 2008; Santostefano et al. 2021).

Fortunately, the estimation can be improved by collecting multiple repeated measurements. If we want to accurately separate the additive genetic, permanent and within-individual variances that could be of interest especially for labile traits, we had to include repeated measurements in the models (Kruuk 2004; Wilson et al. 2010). Furthermore, a previous study found that when the between-individual sample size is large, repeated measurements can lead to more precise heritability estimates (Adams 2014). Although, in our simulation precision increased with the number of measurements in only some specific scenarios (usually when Ve was high and Vpe was low), according to our results, repeated measurements may have other advantages. The quality of heritability estimation of the models containing also repeated measurements depended on the between-individual sample size and on the magnitude of the true heritability as described previously, but collecting 2–5 repeated measurements usually led to 9 and 16% less biased (and in some scenarios more precise, as was mentioned previously) estimation of heritability. Using 10 measurements only offered advantage in some cases (mostly in two scenarios: when Va = Ve = 0.5, Vpe = 0 and when Va = 0.1, Vpe = 0.1 and Ve = 0.8). Overall, the effect of the number of repeated measurements was substantial when Vpe was very low and Ve was high. The effect of repeated measurements in animal models has received little attention, but we can suppose some explanations. If we sample only one measurement from labile traits with high within-individual variability we may obtain biased results (Boake 1989; Dingemanse and Dochtermann 2013; Niemelä and Dingemanse 2018). If the sampled phenotypic values did not reflect well the phenotypic variability of the population, then genetic effects also became difficult to estimate. Thus, more repeated measurements facilitate the less biased and more precise estimation of the residual component reflecting partly the within-individual variance and presumably enables also the reliable separation of additive genetic, permanent environmental and residual effects resulting in good estimation of heritability. The first part of this explanation is corroborated by our results, as we found that the underestimation (or overestimation in some specific cases, see Results and Supplementary Table S5) of heritability was due to the poor partition of Va and Vpe, while Ve was generally reliably estimated with repeated measurements (see Fig. S2). The better separation of the variance components is also probable based on the generally negative trend between number of measurements and relative bias in our results (see Fig. 1h, i).

Nevertheless, our results highlight that the estimation of heritability including repeated measurements in labile traits (when its expected value is low) is not necessarily biased or imprecise, as the relative bias and precision of heritability estimates in the scenarios of Ve = 0.5 or Ve = 0.8 were comparable to the other scenarios in many cases. Additionally, simulation studies for repeatability (which is also a ratio of variance components similarly to heritability) recommend 4 repeated measurements with 100 or 200 individuals that should result in accurate and precise estimates regardless of the value of generating parameter and the complexity of relationships between the variance components (Dingemanse and Dochtermann 2013; Royauté and Dochtermann 2021). Our results generally echo this suggestion, but suggest that in the case of the estimation of heritability sampling 100 individuals may be insufficient even if repeated measurements are taken. However, if the within-individual variance is high and the expected heritability is low, it can be advantageous to collect 2 measurements from 500 individuals than only one measurement from 1000 individuals.

Our results are of special interest for researchers investigating labile traits, such as behaviour, life history or physiological traits. Heritability of behaviour (usually characterized by high within-individual variation) was repeatedly found to be lower (on average 0.30) than that of morphological traits (0.46), while the heritability of life history (0.26) and physiological traits (0.33) was similar (Mousseau and Roff 1987; Stirling et al. 2002). Another review with data from wild populations found on average 0.5 heritability for behavioural traits (Postma 2014). The amount of heritable variation may also depend on whether the behaviour is learnt or not, as for example characteristics of innate calls (0.07 ± 0.05–0.38 ± 0.11, on average 0.21 ± 0.08) had higher heritability than learned song traits (0.03 ± 0.05–0.28 ± 0.09, on average 0.12 ± 0.07) in zebra finches (*Taeniopygia guttata*) (Forstmeier et al. 2009). Furthermore, specific studies on the heritability of behaviour that used multiple measurements from individuals found usually very low values e.g. 0.26 (95% confidence interval (CI): 0.01–0.55) for aggressiveness (2854 test/679 individuals) in great tits (*Parus major*) (Araya-Ajoy and Dingemanse 2017), 0.06 (95% CI: < 0.01–0.17), − 0.10 (95% CI: < 0.01–0.31) for song traits (3582 songs from 81 individuals) in the collared flycatcher (Jablonszky et al. 2022), 0.21 ± 0.07 for locomotor performance (341 tests from 187 individuals) and 0.08 ± 0.04 for vigilance (1237 tests from 315 individuals) in yellow-bellied marmots (*Marmota flaviventris*) (Blumstein et al. 2010) and 0.05 ± 0.02 for aggressiveness (1195 tests from 455 individuals) in Mediterranean field crickets (*Gryllus bimaculatus*) (Santostefano et al. 2017). Regarding life history traits heritability estimates close to 0 ± 0.01 were found in Eastern chipmunks (*Tamias striatus*, 1540 individuals) for fecundity (Santostefano et al. 2021), for clutch size values between 0.15–0.45 were reported for great tits (657–6156 records from 493 to 4077 individuals) and between 0.10 and 0.25 (430–2161 records from 208 to 509 individuals) mute swans (*Cygnus olor*) (Quinn et al. 2006) and 0.11 ± 0.003 for laying date (11,624 observations from 2262 individuals) in common gulls (*Larus canus*) (Brommer et al. 2008). Heritability of various morphological traits (characterized by low within-individual variability) was found between 0.14 ± 0.04—0.42 ± 0.04 (1620–3335 measurements from 720 to 1448 individuals) in house sparrows (*Passer domesticus*), 0.15 ± 0.05–0.29 ± 0.07 (1923–1981 measurements from 790 to 800 individuals) in collared flycatchers (Silva et al. 2017), 0.05 ± 0.10–0.72 ± 0.03 (302–456 individuals) in great reed warblers (*Acrocephalus arundinaceus*) (Åkesson et al. 2008) and 0.26 ± 0.04–0.47 ± 0.07 (2247–2564 measurements from 803 to 891 individuals) in traits of adult sheep (Bérénos et al. 2014) if repeated measurements were included. Thus, many low and non-significant heritability estimates are reported for behavioural and life history traits that underline the importance of our present findings on high relative bias in low heritability estimates even when including 500 or more individuals. Although many of these studies yielded unprecise and non-significant results even with high sample sizes and with multiple measurements, it is still recommended to measure more individuals and collect more repeated measurement as, according to our simulation, these can improve precision and statistical power in some scenarios when within-individual variance is high.

However, it should be noted that repeated measurements did not always improve the goodness of the estimation and in a few cases even decreased precision when the precision of the one measurement models was extremely high (see Fig. 1e, deep blue triangles, but in these cases, precision remained still relatively high with repeated measurements and high precision maybe caused by the Va estimates of the models stuck at zero as Va was 0.1 in these models). In many scenarios, repeated measurements did not have either positive or negative effect on the performance metrics. This may have multiple potential explanations. Despite the large overall sample size, using 100 individuals leads to biased and unprecise heritability estimates; thus, it seems that the repeated measurements could not compensate for the low number of individuals. Bias decreased and power increased with the number of repeated measurements at this small sample size only when the true heritability was high (thus relatively easily estimated; Klein 1974; Charmantier and Réale 2005; Krag et al. 2013)) and the within-individual variance was also high (and permanent environmental effects was low). On the other hand, when using large between-individual sample sizes and the true heritability was high then the estimates were unbiased, so repeated measurements could not offer further improvement at least in terms of bias and power. However, repeated measurement can still improve the estimation even at these large sample sizes when heritability is low and consequently the accuracy and power of estimation is low.

We note that, although we considered 120 scenarios in our study, we could not investigate all potential factors that may influence the accuracy of heritability estimation. Further studies may explore the effect of the relatedness and mistakes in the pedigree (Charmantier and Réale 2005; de Villemereuil et al. 2013; Krag et al. 2013), unequal sampling and various distributions of the response variable on the estimation of heritability (Schielzeth et al. 2020).

In sum, heritability estimates were influenced by the interaction of several factors: the between-individual and within-individual sample sizes, the true value of the additive genetic and within-individual variance. Specifically, heritability can be estimated more precisely and with less bias if 2–10 repeated measurements are taken of the focal trait and this effect can still be significant for higher sample sizes (more than 500 individuals) if the true heritability is low. This advantage is particularly important if the within-individual variance is high, such as in behavioural traits. Thus, we recommend (i) collecting data from more than 100 individuals, (ii) collecting 2–5 repeated measurements and even 10 measurements if within-individual variance is expected to be extremely high when the number of sampled individuals is around 500, and (iii) collecting repeated measurements when the number of individuals is around 1000 only when heritability is expected to be low and within-individual variation is expected to be high).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00265-024-03435-w.

Acknowledgements We thank Pierre de Villemereuil for his statistical advice. We thank Joel Pick and the anonymous reviewers for their helpful comments on a previous version of this manuscript. On behalf of this project, we are grateful for the usage of ELKH Cloud (see (Héder et al. 2022); https://science-cloud.hu/) which helped us achieve the results published in this paper.

Author contribution LZG and MJ designed the study. MJ conducted the statistical analyses with inputs from LZG and wrote the first draft of the manuscript. Both authors contributed to the revision of the manuscript.

Funding Open access funding provided by HUN-REN Centre for Ecological Research. This work was supported by funds from the Hungarian National Research, Development and Innovation Office (K139992).

Data availability The results of the simulations are uploaded as Supplementary files.

Declarations

Ethics approval This study is focusing on a simulation for which no ethical approval is needed.

Conflict of interest The authors declare no competing interests.

References

Adams MJ (2014) Feasibility and uncertainty in behavior genetics for the nonhuman primate. Int J Primatol 35:156–168. https://doi.org/10.1007/s10764-013-9722-8

Åkesson M, Bensch S, Hasselquist D, Tarka M, Hansson B (2008) Estimating heritabilities and genetic correlations: comparing the "animal model" with parent-offspring regression using data from a natural population. PLoS ONE 3:e1739. https://doi.org/10.1371/journal.pone.0001739

Araya-Ajoy YG, Dingemanse NJ (2017) Repeatability, heritability, and age-dependence of seasonal plasticity in aggressiveness in a wild passerine bird. J Anim Ecol 86:227–238. https://doi.org/10.1111/1365-2656.12621

Bell AM, Hankison SJ, Laskowski KL (2009) The repeatability of behaviour: a meta-analysis. Anim Behav 77:771–783. https://doi.org/10.1016/j.anbehav.2008.12.022

Bérénos C, Ellis PA, Pilkington JG, Pemberton JM (2014) Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. Mol Ecol 23:3434–3451. https://doi.org/10.1111/mec.12827

Blumstein DT, Lea AJ, Olson LE, Martin JGA (2010) Heritability of anti-predatory traits: vigilance and locomotor performance in marmots. J Evol Biol 23:879–887. https://doi.org/10.1111/j.1420-9101.2010.01967.x

Boake CRB (1989) Repeatability: its role in evolutionary studies of mating behavior. Evol Ecol 3:173–182. https://doi.org/10.1007/bf02270919

Bourret A, Garant D (2017) An assessment of the reliability of quantitative genetics estimates in study systems with high rate of extra-pair reproduction and low recruitment. Heredity 118:229–238. https://doi.org/10.1038/hdy.2016.92

Brommer JE, Rattiste K, Wilson AJ (2008) Exploring plasticity in the wild: laying date-temperature reaction norms in the common gull *Larus canus*. Proc R Soc Lond B 275:687–693. https://doi.org/10.1098/rspb.2007.0951

Charmantier A, Gienapp P (2014) Climate change and timing of avian breeding and migration: evolutionary versus plastic changes. Evol Appl 7:15–28. https://doi.org/10.1111/eva.12126

Charmantier A, Réale D (2005) How do misassigned paternities affect the estimation of heritability in the wild? Mol Ecol 14:2839–2850. https://doi.org/10.1111/j.1365-294X.2005.02619.x

Clément V, Bibe B, Verrier E, Elsen JM, Manfredi E, Bouix J, Hanocq E (2001) Simulation analysis to test the influence of model adequacy and data structure on the estimation of genetic parameters for traits with direct and maternal effects. Genet Sel Evol 33:369–395. https://doi.org/10.1186/1297-9686-33-4-369

de Villemereuil P, Gimenez O, Doligez B (2013) Comparing parent-offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: a simulation study for

Gaussian and binary traits. Methods Ecol Evol 4:260–275. https://doi.org/10.1111/2041-210x.12011

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

Dohm MR (2002) Repeatability estimates do not always set an upper limit to heritability. Funct Ecol 16:273–280

Fisher RA (1930) The genetical theory of natural selection. The Clarendon Press, Oxford, UK

Forstmeier W, Burger C, Temnow K, Deregnaucourt S (2009) The genetic basis of zebra finch vocalizations. Evolution 63:2114–2130. https://doi.org/10.1111/j.1558-5646.2009.00688.x

Garamszegi LZ (2016) A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons. Anim Behav 120:223–234. https://doi.org/10.1016/j.anbehav.2015.11.009

Ge T, Holmes AJ, Buckner RL, Smoller JW, Sabuncu MR (2017) Heritability analysis with repeat measurements and its application to resting-state functional connectivity. P Natl Acad Sci USA 114:5521–5526. https://doi.org/10.1073/pnas.1700765114

Gorjanc G Henderson DA with code contributions by Kinghorn B, Percy A (2021) GeneticsPed: Pedigree and genetic relationship functions R package version 1.56.0 http://rgenetics.org

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

Hadfield JD, Wilson AJ, Garant D, Sheldon BC, Kruuk LEB (2010) The misuse of BLUP in ecology and evolution. Am Nat 175:116–125. https://doi.org/10.1086/648604

Héder M, Rigó E, Medgyesi D et al (2022) The past, present and future of the ELKH cloud. Inf Társadalom 22:128. https://doi.org/10.22503/inftars.xxii.2022.2.8

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

Israel C, Weller JI (2000) Effect of misidentification on genetic gain and estimation of breeding value in dairy cattle populations. J Dairy Sci 83:181–187. https://doi.org/10.3168/jds.S0022-0302(00)74869-7

Jablonszky M, Canal D, Hegyi G et al (2022) Estimating heritability of song considering within-individual variance in a wild songbird: the collared flycatcher. Front Ecol Evol 10:975687. https://doi.org/10.3389/fevo.2022.975687

Klein TW (1974) Heritability and genetic correlation: statistical power, population comparisons, and sample size. Behav Genet 4:171–189. https://doi.org/10.1007/bf01065758

Kominakis AP (2008) Effect of unfavourable population structure on estimates of heritability, systematic effects and breeding values. Arch Tierzucht 51:601–610. https://doi.org/10.5194/aab-51-601-2008

Krag K, Janss LL, Shariati MM, Berg P, Buitenhuis AJ (2013) SNP-based heritability estimation using a Bayesian approach. Animal 7:531–539. https://doi.org/10.1017/s1751731112002017

Kruuk LEB (2004) Estimating genetic parameters in natural populations using the "animal model." Phil Trans R Soc B 359:873–890. https://doi.org/10.1098/rstb.2003.1437

Kruuk LEB, Hadfield JD (2007) How to separate genetic and environmental causes of similarity between relatives. J Evol Biol 20:1890–1903. https://doi.org/10.1111/j.1420-9101.2007.01377.x

Liang LM, Chen WM, Sham PC, Abecasis GR (2009) Variance components linkage analysis with repeated measurements. Hum Hered 67:237–247. https://doi.org/10.1159/000194977

Martin JGA, 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. https://doi.org/10.1111/j.2041-210X.2010.00084.x

Morrissey MB, Wilson AJ, Pemberton JM, Ferguson MM (2007) A framework for power and sensitivity analyses for quantitative genetic studies of natural populations, and case studies in Soay sheep (*Ovis aries*). J Evol Biol 20:2309–2321. https://doi.org/10.1111/j.1420-9101.2007.01412.x

Mousseau TA, Roff DA (1987) Natural-selection and the heritability of fitness components. Heredity 59:181–197. https://doi.org/10.1038/hdy.1987.113

Niemelä PT, Dingemanse NJ (2018) On the usage of single measurements in behavioural ecology research on individual differences. Anim Behav 145:99–105. https://doi.org/10.1016/j.anbehav.2018.09.012

Pick J, Kasper C, Allegue H et al (2022) Describing posterior distributions of variance components: problems and the use of null distributions to aid interpretation. EcoEvoRxiv. https://doi.org/10.1111/2041-210X.14200

Piersma T, Drent J (2003) Phenotypic flexibility and the evolution of organismal design. Trends Ecol Evol 18:228–233. https://doi.org/10.1016/S0169-5347(03)00036-3

Postma E (2014) Four decades of estimating heritabilities in wild vertebrate populations: improved methods, more data, better estimates? In: Charmantier A, Garant D, Kruuk LEB (eds) Quantitative Genetics in the Wild. Oxford University Press, Oxford, pp 16–33

Quinn JL, Charmantier A, Garant D, Sheldon BC (2006) Data depth, data completeness, and their influence on quantitative genetic estimation in two contrasting bird populations. J Evol Biol 19:994–1002. https://doi.org/10.1111/j.1420-9101.2006.01081.x

R Core Team (2019) R: A language and environment for statistical computing R Foundation for Statistical Computing Vienna Austria http://www.R-project.org

Raffa JD, Thompson EA (2016) Power and effective study size in heritability studies. Stat Biosci 8:264–283. https://doi.org/10.1007/s12561-016-9143-2

Risk BB, Zhu HT (2018) Note on bias from averaging repeated measurements in heritability studies. P Natl Acad Sci USA 115:E122–E122. https://doi.org/10.1073/pnas.1719250115

Royauté R, Dochtermann NA (2021) Comparing ecological and evolutionary variability within datasets. Behav Ecol Sociobiol 75:127. https://doi.org/10.1007/s00265-021-03068-3

Santostefano F, Allegue H, Garant D, Bergeron P, Réale D (2021) Indirect genetic and environmental effects on behaviors, morphology, and life-history traits in a wild Eastern chipmunk population. Evolution 75:1492–1512. https://doi.org/10.1111/evo.14232

Santostefano F, Wilson AJ, Niemelä PT, Dingemanse NJ (2017) Indirect genetic effects: a key component of the genetic architecture of behaviour. Sci Rep 7:10235. https://doi.org/10.1038/s41598-017-08258-6

Schielzeth H, Dingemanse NJ, Nakagawa S, Westneat DF, Allegue H, Teplitsky C, Réale D, Dochtermann NA, Garamszegi LZ, Araya-Ajoy YG (2020) Robustness of linear mixed-effects models to violations of distributional assumptions. Methods Ecol Evol 11:1141–1152. https://doi.org/10.1111/2041-210x.13434

Schoenemann KL, Bonier F (2018) Repeatability of glucocorticoid hormones in vertebrates: a meta-analysis. PeerJ 6:e4398. https://doi.org/10.7717/peerj.4398

Silva CNS, McFarlane SE, Hagen IJ et al (2017) Insights into the genetic architecture of morphological traits in two passerine bird species. Heredity 119:197–205. https://doi.org/10.1038/hdy.2017.29

Snell-Rood EC (2013) An overview of the evolutionary causes and consequences of behavioural plasticity. Anim Behav 85:1004–1011. https://doi.org/10.1016/j.anbehav.2012.12.031

Steinsland I, Larsen CT, Roulin A, Jensen H (2014) Quantitative genetic modeling and inference in the presence of nonignorable missing data. Evolution 68:1735–1747. https://doi.org/10.1111/evo.12380

Stirling DG, Réale D, Roff DA (2002) Selection, structure and the heritability of behaviour. J Evol Biol 15:277–289. https://doi.org/10.1046/j.1420-9101.2002.00389.x

Taff CC, Schoenle LA, Vitousek MN (2018) The repeatability of glucocorticoids: a review and meta-analysis. Gen Comp Endocrinol 260:136–145. https://doi.org/10.1016/j.ygcen.2018.01.011

Weigensberg I, Roff DA (1996) Natural heritabilities: can they be reliably estimated in the laboratory? Evolution 50:2149–2157. https://doi.org/10.2307/2410686

Westneat DF, Araya-Ajoy YG, Allegue H et al (2020) Collision between biological process and statistical analysis revealed by mean centring. J Anim Ecol 89:2813–2824. https://doi.org/10.1111/1365-2656.13360

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

Zhang HP, Zhong XY (2006) Linkage analysis of longitudinal data and design consideration. BMC Genet 7:37. https://doi.org/10.1186/1471-2156-7-37

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.