Heritability, selection, and the response to selection in the presence of phenotypic measurement error: effects, cures, and the role of repeated measurements

Erica Ponzi[1,2], Lukas F. Keller[1,3], Timothée Bonnet[1,4], Stefanie Muff[1,2,⋆]

Running title: Phenotypic measurement error – effects and cures

[1] Institute of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland

[2] Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zürich, Hirschengraben 84, 8001 Zürich, Switzerland

[3] Zoological Museum, University of Zürich, Karl-Schmid-Strasse 4, 8006 Zürich, Switzerland

[4] Division of Ecology and Evolution, Research School of Biology,The Australian National University, Acton, Canberra, ACT 2601,Australia

⋆ Corresponding author. E-mail: stefanie.muff@uzh.ch

Quantitative genetic analyses require extensive measurements of phenotypic traits, which may be especially challenging to obtain in wild populations. On top of operational measurement challenges, some traits undergo transient fluctuations that might be irrelevant for selection processes. The presence of transient variability, denoted here as measurement error in a broader sense, is a possible cause for bias in the estimation of quantitative genetic parameters. We illustrate how transient effects with a classical measurement error structure may bias estimates of heritability, selection gradients, and the response to selection of a continuous trait, and propose suitable strategies to obtain error-corrected estimates. Unbiased estimates can be retrieved with repeated measurements, given that they were taken at an appropriate temporal scale. However, the fact that repeated measurements were previously used to isolate permanent environmental (instead of transient) effects requires that we rethink their information content. We propose a distinction into "short-term" and "long-term" repeats, whereas the former capture transient variability, and the latter the permanent effects. A simulation study illustrates how the inclusion of appropriate variance components in the models helps retrieving unbiased estimates of all quantities of interest. The methodology is then applied to data from a Swiss snow vole population.

Keywords: Measurement error, animal model, breeder's equation, Robertson-Price identity, permanent environmental effects, error variance.

Introduction

Quantitative genetic methods are central to understand and predict evolutionary change of phenotypic traits (Falconer and Mackay, 1996; Lynch and Walsh, 1998; Charmantier et al., 2014). At its core, quantitative genetics is a statistical approach that decomposes the observed phenotype P into the sum of additive genetic effects A and environmental effects E, so that $P = A+E$. This basic model can be extended in various ways (Falconer and Mackay, 1996; Lynch and Walsh, 1998), with one of the most common being $P = A + PE + E$, where PE captures the so-called *permanent environmental effects*. Permanent environmental effects are stable differences among individuals that are induced by the environment, and these can be isolated when repeated measurements of an individual covary more than expected from the genetic effects alone (Lynch and Walsh, 1998; Kruuk, 2004; Wilson et al., 2010).

This quantitative genetic decomposition of phenotypes is not possible at the individual level, but under the crucial assumption of independence of genetic, permanent environmental, and environmental effects, the phenotypic variance at the population level can be decomposed as $\sigma_P^2 = \sigma_A^2+\sigma_{PE}^2+\sigma_E^2$. The additive genetic variance (σ_A^2) is often the center of interest in quantitative genetics, since the amount of evolutionary change in a phenotypic trait under selection is a function of σ_A^2 under both commonly used frameworks to quantify the response to selection, the *breeder's equation* (BE) and the *Robertson-Price identity*, also known as *Robertson's secondary theorem of selection* (STS). The breeder's equation predicts the response to selection of a trait $\boldsymbol{z}$ from the product of the heritability of the trait and the strength of selection

$$R_{\mathrm{BE}} = h^2 \cdot S \tag{1}$$

(Lush, 1937; Falconer and Mackay, 1996), where R_{BE} is the response to selection for trait $\boldsymbol{z}$, S is the selection differential, defined as the mean phenotypic difference between selected individuals and the population mean, and h^2 is the proportion of additive genetic to total phenotypic variance

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \ . \tag{2}$$

Under the secondary theorem of selection, evolutionary change is equal to the additive genetic covariance of a trait with relative fitness ($\boldsymbol{w}$), that is,

$$R_{\mathrm{STS}} = \sigma_a(\boldsymbol{z}, \boldsymbol{w}) \tag{3}$$

(Robertson, 1966; Price, 1970). Morrissey et al. (2010) and Morrissey et al. (2012) discuss the differences between the breeder's equation and the secondary theorem of

selection in detail.

The reliable estimation of the (co-)variance parameters of interest requires large amounts of data, often collected across multiple generations, with known relationships among individuals in the data set. For many phenotypic traits of interest, data collection is often not trivial, and multiple sources of error, such as phenotypic measurement error, pedigree errors (wrong relationships among individuals), or non-randomly missing data may affect the parameter estimates. Pedigree errors and problems arising from missing data have been discussed and addressed in a number of studies (*e.g.* Keller et al., 2001; Griffith et al., 2002; Senneke et al., 2004; Charmantier and Reale, 2005; Steinsland et al., 2014; Ponzi et al., 2017; Wolak and Reid, 2017). In contrast, although known for a long time (*e.g.* Price and Boag, 1987) the effects of phenotypic measurement error on estimates of (co-)variance components have received less attention (but see *e.g.* Dohm, 2002; Macgregor et al., 2006; van der Sluis et al., 2010; Ge et al., 2017). In particular, general solutions to obtaining unbiased estimates of (co-)variance parameters in the presence of phenotypic measurement error are lacking.

In the simplest case, and the case considered here, phenotypic measurement error is assumed to be independent and additive, that is, instead of the actual phenotype P, an error-prone version $P + e$ is measured, where e denotes an error term with variance $\sigma^2_{e_m}$ (see p.121 Lynch and Walsh, 1998). As a consequence, the total phenotypic variance of the recorded values is $\sigma^2_P + \sigma^2_{e_m}$ instead of σ^2_P. Consequently, measurement error in the phenotype leads to an observed phenotypic variance that is larger than its actual value, from which the error variance $\sigma^2_{e_m}$ should therefore be disentangled in order to estimate unbiased parameters. However, most existing methods for continuous trait analyses that acknowledged measurement error have modeled it as part of the environmental component, and thus implicitly as part of the total phenotypic value (*e.g.* Dohm, 2002; Macgregor et al., 2006; van der Sluis et al., 2010). This means that in the decomposition of a phenotype $P = A + PE + E$, measurement error is absorbed in E, thus $\sigma^2_{e_m}$ is absorbed by σ^2_E. Under the assumption taken here that measurement error is independent, it does not affect the relatedness among individuals and therefore leaves the estimate of the additive genetic variance σ^2_A unaffected. This practice effectively *downwardly* biases heritability estimates, because

$$h^2_{\text{biased}} = \frac{\sigma^2_A}{\sigma^2_P + \sigma^2_{e_m}} \leq \frac{\sigma^2_A}{\sigma^2_P} , \tag{4}$$

see for example Lynch and Walsh (p.121, 1998) or Ge et al. (2017).

To obtain a correct estimate of h^2 it is thus necessary to disentangle $\sigma^2_{e_m}$ from the actual phenotypic variance σ^2_P, and particularly from its environmental component σ^2_E. Ge et al. (2017) suggested to isolate $\sigma^2_{e_m}$ by taking repeated phenotypic mea-

surements of the same individual. Ge et al. (2017) interpret measurement error in a broad sense as "transient effects", that is, fluctuations not only due to operational measurement imprecision, but also due to short-term changes of the phenotype itself that are not relevant for the selection process (see below for an example). This use of repeated measurements to separate *transient* fluctuations, such as measurement error variance ($\sigma^2_{e_m}$), from the environmental variance (σ^2_E) contrasts with the usual use of repeated measures in quantitative genetics, which is to separate the variance of *stable* permanent environmental effects (σ^2_{PE}) from the additive genetic variance (σ^2_A)(*e.g.* Kruuk and Hadfield, 2007; Logan et al., 2016; Notter et al., 2017). In the first case, repeated measures are used to prevent *downward* bias in estimates of h^2, in the second to prevent *upward* bias in σ^2_A and h^2 (Wilson et al., 2010).

The key difference between these two uses of repeated phenotypic measurements is that measurement error is regarded as undesirable "noise" that captures transient effects not actually "seen" by the selection process (see *e.g.* Price and Boag, 1987, p. 279 for a similar analogy), while the permanent environmental effect captures trait characteristics that are persistent (non-transient) for an individual and are thus potentially under selection. As an example, if the trait is the mass of an adult animal, repeated measurements within the same day are expected to differ even in the absence of instrumental error, simply because animals eat, drink and defecate (for an example see Keller and Van Noordwijk, 1993). Such transient (short-term) effects might not actually be of interest if there is reason to believe that they do not contribute to the selection process. Therefore, they can also be regarded as measurement error in a broader sense – although, strictly speaking, they are not actual mismeasurements, but rather fluctuations in the trait that one might want to exclude from the analysis.

The aim of this article is to develop general methods to obtain unbiased estimates of heritability, selection, and response to selection in the presence of measurement error, building on the work by Ge et al. (2017). We start by clarifying the meaning and information content of repeated phenotypic measurements on the same individual. The type of phenotypic trait we have in mind is a relatively plastic trait, such as milk production or an animal's mass, which are expected to undergo changes across an individual's lifespan that are biologically relevant. We show that repeated measures taken over different time intervals can help separate transient effects from more stable (permanent) environmental and genetic effects. We proceed to show that based on such a variance decomposition one can construct models that yield unbiased estimates of heritability, selection, and the response to selection in the presence of phenotypic measurement error. We illustrate these approaches with a simulation study, and with empirical quantitative trait analyses of body mass measurements taken in a population of snow voles in the Swiss alps (Bonnet et al., 2017). Finally,

we highlight another crucial advantage of the secondary theorem of selection over the breeder's equation, namely that the genetic covariance between a focal trait and relative fitness is robust against independent phenotypic measurement error because covariances are not affected if the error is "randomized over phenotypes", that is, if it is independent of the actual trait value (Price and Boag, 1987, p. 277).

Material and methods

Heritability

Estimating heritability for error-free phenotypes

In the absence of phenotypic measurement error of a focal trait $\boldsymbol{z}$, heritability is estimated as indicated in equation (2). In the presence of repeated measurements j of individuals i taken at different life stages, the variance components of the decomposition $\sigma_p^2(\boldsymbol{z}) = \sigma_A^2 + \sigma_{PE}^2 + \sigma_E^2$ are typically estimated via an *animal model* (Henderson, 1976; Lynch and Walsh, 1998; Kruuk, 2004), a (generalized) linear mixed model that decomposes the measurements of the trait z_{ij} and accounts for repeats as

$$z_{ij} = \mu + \boldsymbol{\beta}^\top \boldsymbol{X}_i + a_i + id_i + E_{ij} \ , \tag{5}$$

where μ is the population mean, $\boldsymbol{\beta}$ is a vector of fixed effects and $\boldsymbol{X}_i$ is the vector of covariates for animal i. The remaining components are the random effects, namely a_i, the breeding value with dependency structure $(a_1, \ldots, a_n)^T \sim \mathsf{N}(0, \sigma_A^2 \mathbf{A})$, an independent, animal-specific permanent environmental effect $id_i \sim \mathsf{N}(0, \sigma_{PE}^2)$, and an independent Gaussian residual term $E_{ij} \sim \mathsf{N}(0, \sigma_E^2)$ that captures the remaining (unexplained) variability. The dependency structure of the breeding values a_i is encoded by the additive genetic relatedness matrix $\mathbf{A}$, which was traditionally derived from a pedigree, but can alternatively be calculated from genomic data (Meuwissen et al., 2001; Hill, 2014). Without inbreeding, the entry (l, m) of $\mathbf{A}$ is given by $2 \cdot \theta_{lm}$, where θ_{lm} is the probability that a gene copy drawn at random from individual l will be identical by descent to a gene copy drawn at random from individual m, known as the coefficient of co-ancestry and equal to 1 on the diagonal entries of the matrix. In the presence of inbreeding, the diagonal entry (l, l) is given by $1 + F_l$, where F_l is the inbreeding coefficient of individual l (see Lynch and Walsh (1998), p.77). Model (5) can be further expanded to include more fixed or random effects, such as maternal, nest or time effects.

Estimating heritability with phenotypic measurement error

Now assume that the actual values z_{ij} cannot directly be measured, but instead an error-prone proxy $z^{\star}_{ij}$ that obscures z_{ij} according to a *classical measurement error model* (Carroll et al., 2006)

$$z^{\star}_{ij} = z_{ij} + e_{ij} \quad \text{with} \quad e_{ij} \sim \mathsf{N}(0, \sigma^2_{e_m}) \ , \tag{6}$$

where $\sigma^2_{e_m}$ is the error variance of the measurements and the e_{ij} are assumed independent of each other, of z_{ij}, and of any other variables. The total *observed* phenotypic variance is then $\sigma^2_p(\boldsymbol{z}) + \sigma^2_{e_m}$, thus larger than the *actual* phenotypic variance. If such data are analyzed via model (5), the error variance $\sigma^2_{e_m}$ is absorbed by the environmental variance σ^2_E. In the presence of measurement error, we will denote model (5) as the *naive* model, which leads to an overestimated phenotypic variance and, in particular, to a biased version $h^2_{\text{biased}} \leq h^2$ as given in equation (4).

Rethinking repeated measurements

To understand the role of repeated measurements and the variance components they can isolate, it is crucial to distinguish among repeats that are taken within short-term intervals (or even at the same time point), and repeats taken at long-term distance, such as monthly sessions, seasons or years. We introduce the notion of a *measurement session* for short-term intervals within which the phenotype of a focal trait is not expected to undergo phenotypic changes that are relevant for selection, which facilitates the discrimination of within-session (short-term) and between-session (long-term) repeats. In the introduction we employed the example of an animal's mass that transiently fluctuates within a day. Depending on the context, such fluctuations might not be of interest, and the "actual" phenotypic value would correspond to the average daily mass. A measurement session could then for instance correspond to a single day, and within-session repeats help to separate σ^2_E and $\sigma^2_{e_m}$. In the absence of long-term repeats, the model that correctly mirrors the data structure includes an environmental effect $E_i \sim \mathsf{N}(0, \sigma^2_E)$ and an error term $e_{ik} \sim \mathsf{N}(0, \sigma^2_{e_m})$:

$$z_{ik} = \mu + \boldsymbol{\beta}^{\top} \boldsymbol{X}_i + a_i + E_i + e_{ik} \ . \tag{7}$$

The situation is depicted in Figure 1, case a). On the other hand, when measurements are taken at long-term distances, that is, in different measurement sessions (*e.g.* across seasons or years), it is no longer assumed that the actual trait value remains constant. Therefore, such repeats are able to separate the permanent environmental from stochastic (changing) environmental effects, and they are the types of repeats that are traditionally used by quantitative geneticists to disentangle σ^2_{PE}

from σ_A^2. The situation is depicted in Figure 1, case b), and the corresponding animal model is then given as

$$z_{ij} = \mu + \boldsymbol{\beta}^\top \boldsymbol{X}_i + a_i + id_i + E_{ij} \ , \tag{8}$$

with permanent environmental effect $id_i \sim \mathsf{N}(0, \sigma_{PE}^2)$ and an environmental term $E_{ij} \sim \mathsf{N}(0, \sigma_E^2)$. Finally, if *both* close and distant repeats are present, that is, when individual i is measured in different sessions j and with repeats k within sessions, it is possible to separate all three variance components (Figure 1, case c)) via model

$$z_{ijk} = \mu + \boldsymbol{\beta}^\top \boldsymbol{X}_i + a_i + id_i + E_{ij} + e_{ijk} \ , \tag{9}$$

with $id_i \sim \mathsf{N}(0, \sigma_{PE}^2)$, $E_{ij} \sim \mathsf{N}(0, \sigma_E^2)$ and $e_{ijk} \sim \mathsf{N}(0, \sigma_{e_m}^2)$.

Selection

Selection occurs when a trait is correlated with fitness, such that variations in the trait values lead to predictable variations among the same individuals in fitness. The leading approach for measuring the strength of directional selection is the one developed by Lande and Arnold (1983), who proposed to estimate the selection gradient β_z as the slope of the regression of relative fitness $\boldsymbol{w}$ on the phenotypic trait $\boldsymbol{z}$

$$\boldsymbol{w} = \boldsymbol{\alpha} + \beta_z \cdot \boldsymbol{z} + \boldsymbol{\epsilon} \ , \tag{10}$$

with intercept $\boldsymbol{\alpha}$ and residual error vector $\boldsymbol{\epsilon}$. This model can be further extended to correct for covariates, such as sex or age. If the phenotype is measured with classical measurement error, such that the observed value is $\boldsymbol{z}^\star = \boldsymbol{z} + \boldsymbol{e}$ with error variance $\sigma_{e_m}^2$ as in (6) (where vector notation is used here for notational convenience), the regression of $\boldsymbol{w}$ against $\boldsymbol{z}^\star$ leads to an attenuated version of β_z (Mitchell-Olds and Shaw, 1987; Fuller, 1987; Carroll et al., 2006). Using that $\hat{\beta}_z = \frac{\sigma(\boldsymbol{z},\boldsymbol{w})}{\sigma_p^2(\boldsymbol{z})}$, $\sigma_p^2(\boldsymbol{z}^\star) = \sigma_p^2(\boldsymbol{z}) + \sigma_{e_m}^2$, and the assumption that the error in $\boldsymbol{z}^\star$ is independent of $\boldsymbol{w}$, simple calculations show that the error-prone estimate of selection will in fact be given by

$$\hat{\beta}_z^\star = \frac{\sigma(\boldsymbol{z}^\star, \boldsymbol{w})}{\sigma_p^2(\boldsymbol{z}^\star)} = \frac{\sigma(\boldsymbol{z}, \boldsymbol{w})}{\sigma_p^2(\boldsymbol{z}) + \sigma_{e_m}^2} \leq \hat{\beta}_z \ ,$$

and equality holds only in the absence of error, that is, if $\sigma_{e_m}^2 = 0$. To obtain an unbiased estimate of selection it may thus often be necessary to correct for the error by a suitable error-correction method. Such a correction must rely on the same type of short-term repeated measurements that allow to isolate transient effects in the animal model, as those given *e.g.* in equations (7) or (9).

To include such repeats in model (10), we rely on the interpretation as an error-in-variables problem and formulate it as a Bayesian hierarchical model, where low-level units are organized into a hierarchy of successively higher-level units. It is relatively straightforward to explicitly add a model for the error by specifying an additional level for the error-prone covariate. This formulation, together with the possibility to include prior knowledge, provides a flexible way to model measurement error (Stephens and Dellaportas, 1992; Richardson and Gilks, 1993). To account for error in selection gradients, we formulate a three-level hierarchical model. The first level is the regression model for selection, and the second level is given by the error model of the observed covariate $\boldsymbol{z}^\star$ given its true value $\boldsymbol{z}$. Third, a so-called *exposure model* for the unobserved (true) trait value is required to inform the model about the distribution of $\boldsymbol{z}$, and it seemed natural to employ the animal model (5) for this purpose. Using the notation previously employed for an individual i, measured in different sessions j and with repeats k within sessions, the formulation of the three-level hierarchical model is given as

$$w_{ij} = \alpha + \beta_z z_{ij} + \boldsymbol{\beta}^\top \boldsymbol{X}_i + \epsilon_{ij} \ , \qquad \epsilon_{ij} \sim \mathsf{N}(0, \sigma^2_\epsilon) \qquad \text{Selection model} \qquad (11)$$

$$z^\star_{ijk} = z_{ij} + e_{ijk} \ , \qquad e_{ijk} \sim \mathsf{N}(0, \sigma^2_{e_m}) \qquad \text{Error model} \qquad (12)$$

$$z_{ij} = \mu + \boldsymbol{\beta}^\top \boldsymbol{X}_i + a_i + id_i + E_{ij} \ , \qquad E_{ij} \sim \mathsf{N}(0, \sigma^2_E) \qquad \text{Exposure model} \qquad (13)$$

where w_{ij} is the measurement of the relative fitness trait for individual i, usually taken only once per individual and having the same value for all sessions j, $\boldsymbol{\beta}$ is a vector of fixed effects, $\boldsymbol{X}_i$ is the vector of covariates for animal i , β_z is the estimate of selection and α and ϵ_{ij} are respectively the intercept and the residual term from the linear regression model. The classical measurement error term is given by e_{ijk}. This formulation as a hierarchical model allows to estimate the selection gradient β_z for the error-free covariate $\boldsymbol{z}$, because the lower levels of the model properly account for the error by explicitly describing it. It might be helpful to see that the second and third levels are just a hierarchical representation of model (9). The model can be fitted in a Bayesian setup, see for instance Muff et al. (2015) for a description of how to implement such models in INLA (Rue et al., 2009) via its `R` interface `R-INLA`.

Note that model (11) is formulated here for directional selection. Although the explicit discussion of alternative selection mechanisms, such as stabilizing or disruptive selection, is out of the scope of the present paper, error correction for these cases is straightforward, given that the error model is the same. The only change is that the linear selection model (11) is replaced by the appropriate alternative that *e.g.* includes higher order terms (*e.g.* Fisher, 1930, p. 105).

Response to selection

The breeder's equation

Evolutionary response to a selection process of a phenotypic trait that is related to fitness can be estimated either by the breeder's equation (1) or by the Robertson-Price identity (3), and these two approaches are equivalent only when the respective trait value (in the univariate model) is the sole causal factor for fitness (Morrissey et al., 2010, 2012). Lande (1979) proposed to generalize the breeder's equation to multiple traits, but the implicit assumption then still is that *all* correlated traits causally related to fitness are included in the model. Given that fitness is a complex trait that usually depends on many unmeasured variables (Møller and Jennions, 2002; Peek et al., 2003), it is not surprising that the breeder's equation is often not successful in predicting evolutionary change in natural systems (Hadfield, 2008; Morrissey et al., 2010), in contrast to (artificial) animal breeding situations, where, thanks to the large control on the process, all the traits affecting fitness are known and included in the models (Lush, 1937; Falconer and Mackay, 1996; Roff, 2007).

In the presence of phenotypic measurement error of the type given in equation (6), another disadvantage of the breeder's equation becomes apparent: If the error is not accounted for, the bias in h^2_{biased} directly propagates to the estimated response to selection via $R_{\text{biased}} = h^2_{\text{biased}} \cdot S$. Note that the bias is not counterbalanced by the selection differential S, because for independent error, $\hat{\sigma}_p(\boldsymbol{z}^\star, \boldsymbol{w})$ is an unbiased estimator for $S = \sigma_p(\boldsymbol{z}, \boldsymbol{w})$.

The Robertson-Price identity

The secondary theorem of selection is more robust than the breeder's equation and makes fewer assumptions, in particular with regard to causal relations between the focal trait(s) and fitness (Morrissey et al., 2010, 2012). The genetic covariance of the relative fitness $\boldsymbol{w}$ and the phenotypic trait $\boldsymbol{z}$, $\sigma_a(\boldsymbol{w}, \boldsymbol{z})$ can be estimated from a multivariate animal model. If the evolutionary response of a single trait is to be estimated, the model for the response vector including the trait values $\boldsymbol{z}$ and relative fitness values $\boldsymbol{w}$ is bivariate with

$$\begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{w} \end{bmatrix} = \boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{a} + \mathbf{Z}\boldsymbol{r} \ , \tag{14}$$

where $\boldsymbol{\mu}$ is the intercept vector, $\boldsymbol{\beta}$ the vector of fixed effects, $\boldsymbol{X}$ the corresponding design matrix, $\mathbf{D}$ is the design matrix for the breeding values $\boldsymbol{a}$, and $\mathbf{Z}$ is a design matrix for additional random terms $\boldsymbol{r}$. These may include environmental and/or error terms, depending on the structure of the data that may correspond to the

univariate cases of equations (7) - (9), or again to other random terms such as maternal or nest effects. The actual component of interest is the vector of breeding values, which is assumed multivariate normally distributed with

$$\boldsymbol{a} = \begin{bmatrix} \boldsymbol{a}(z) \\ \boldsymbol{a}(w) \end{bmatrix} \sim \mathsf{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_a^2(\boldsymbol{z})\mathbf{A} & \sigma_a(\boldsymbol{w},\boldsymbol{z})\mathbf{A} \\ \sigma_a(\boldsymbol{w},\boldsymbol{z})\mathbf{A} & \sigma_a^2(\boldsymbol{w})\mathbf{A} \end{bmatrix}\right) , \tag{15}$$

where $\boldsymbol{a}(z)$ and $\boldsymbol{a}(w)$ are the respective subvectors for the trait and fitness, and $\mathbf{A}$ is the relationship matrix derived from the pedigree. An estimate of the additive genetic covariance $\sigma_a(\boldsymbol{w},\boldsymbol{z})$ is extracted from this covariance matrix.

An interesting feature of the respective covariance is that it is unaffected by the presence of independent error in the phenotype, namely because the error variance is captured by the non-genetic variance components $\boldsymbol{r}$ in equation (14) (see *e.g.* p.121 Lynch and Walsh, 1998, where this aspect is discussed for univariate models). This theoretical robustness of R_{STS} against measurement error in the phenotype is certainly another advantage over the breeder's equation, although it depends crucially on the independence assumption of the error.

Simulation study

Generating simulated data

We carried out a simulation study to quantify the expected effect of error in the phenotype on the estimates of heritability, selection and response to selection. Pedigree structures including 6 generations with 100 individuals each (*i.e.* $N = 600$ individuals) were generated using the function `generatePedigree` from the `R` package `pedigree` (Coster, 2013; R Core Team, 2017). The Ne/Nc ratio, *i.e.* the proportion of the effective population size (Ne) with respect to its census size (Nc), was fixed to 1, which was obtained by setting the number of fathers and mothers per generation to 50, such that the number of offsprings per individual was on average equal to 2 and ranged from 1 to 6. This value of Ne/Nc characterizes a rather dense pedigree structure (Palstra and Fraser, 2012), but heritability, selection and response to selection do not change with pedigree density, since they are solely determined by variance and covariance components, and by the trait values. Despite this, the simulations were also repeated with Ne/Nc ratios of 0.50 and 0.30, but since the results did not change qualitatively, they are not shown here.

Error-free phenotypes z_{ij} for individuals $i = 1, ..., 600$ from sessions (*i.e.* at long-term distance) $j = 1, ..., 4$ and one fitness trait w_i per individual were generated

from models with purely additive effects

$$
\begin{aligned}
z_{ij} &= \mu_z + \beta_{\text{sex}} sex_i + a(z)_i + id(z)_i + E_{ij} \\
w_i &= \mu_w + a(w)_i + id(w)_i. \qquad (16)
\end{aligned}
$$

Data were generated such that the variance components summed up to 1, with a true value of heritability equal to 0.3. More specifically,

- breeding values $(\boldsymbol{a}(w), \boldsymbol{a}(z))$ were generated according to equation (15), sampled with the function `rbv` from the `R` package `MCMCglmm` (Hadfield, 2010). The respective variances were set to $\sigma_a^2(\boldsymbol{w}) = 0.3$, $\sigma_a^2(\boldsymbol{z}) = 0.3$ and the genetic correlation between the traits was set to $\sigma_a(\boldsymbol{w}, \boldsymbol{z}) = 0.15$, which corresponds to the value of response to selection according to the Robertson-Price equation.

- the independent permanent environmental effects $(\boldsymbol{id}(w), \boldsymbol{id}(z))$ were generated jointly normal for the two traits using $\sigma_{PE}^2(\boldsymbol{w}) = 0.3$, $\sigma_{PE}^2(\boldsymbol{z}) = 0.3$ and $\sigma_{PE}(\boldsymbol{w}, \boldsymbol{z}) = 0.15$.

- the independent environmental effect was generated for the phenotypic trait according to $E_{ij} \sim \mathsf{N}(0, \sigma_E^2)$ with $\sigma_E^2 = 0.4$. This effect remains constant within a session j of individual i, that is, the repeats are assumed close enough in time to encounter a stable environment. No such component was generated for the fitness trait, as indicated by (16).

- population means were $\mu_z = 10$ and $\mu_w = 1$, and the sex-effect on the phenotypic trait was $\beta_{\text{sex}} = 2$, while there was no sex difference in fitness.

We then obscured the actual phenotypic values z_{ij} by adding measurement error for repeated measurements $k = 1, \ldots, 4$ that were generated within each session. Error in the phenotype was added according to equation

$$
z_{ijk}^{\star} = z_{ij} + e_{ijk} \ ,
$$

with independent measurement error $e_{ijk} \sim \mathsf{N}(0, \sigma_{e_m}^2)$. The full simulation study was repeated with three different error variances $\sigma_{e_m}^2 = 0.15$, 0.25 or 0.50. Fitness traits $\boldsymbol{w}$ were assumed to be error-free. For each of these cases, a total of 100 iterations were simulated.

Analysis of simulated data

For each estimate of interest, three models were fitted to each of the 100 pedigrees: a first model involving the error-free trait $\boldsymbol{z}$ (denoted as the "actual" model), and two models including the error-prone trait $\boldsymbol{z}^{\star}$, once without ("naive") and once with

error modeling ("corrected"). All analyses were performed with `MCMCglmm`, except for the corrected model for selection, which was implemented with `INLA`. The respective `R` script for simulations and analysis of simulated data is reported in Appendix 2.

Heritability. The actual values for heritability were estimated by fitting an animal model with the error-free response z_{ij}, sex as fixed effect and random effects given by the permanent genetic (a_i), permanent environmental (id_i) and environmental (E_{ij}) effects. Because in the error-free case no repeats were available within a session j the respective model equation is given as

$$z_{ij} = \mu + \beta_{\text{sex}} sex_i + a_i + id_i + E_{ij} \quad \text{(actual)} \ , \tag{17}$$

with $(a_1, \ldots, a_n)^\top \sim \mathsf{N}(0, \sigma_A^2 \mathbf{A})$, $id_i \sim \mathsf{N}(0, \sigma_{PE}^2)$ and $E_{ij} \sim \mathsf{N}(0, \sigma_E^2)$. An estimate of heritability was computed according to equation (2). When the phenotypic trait was affected by error and repeated measurements $z^\star_{ijk}$ were available as a proxy to z_{ij}, we formulated two models. First, a naive model which assumes that all variability which is not captured by additive genetic or permanent environmental effects is coming from the environment

$$z^\star_{ijk} = \mu + \beta_{\text{sex}} sex_i + a_i + id_i + E_{ijk} \quad \text{(naive)} \ , \tag{18}$$

with environmental component $E_{ijk} \sim \mathsf{N}(0, \sigma_E^2)$. And second, a corrected model that properly separates the environmental component E_{ij} from the measurement error e_{ijk} by using repeated measurements $z^\star_{ijk}$

$$z^\star_{ijk} = \mu + \beta_{\text{sex}} sex_i + a_i + id_i + E_{ij} + e_{ijk} \quad \text{(corrected)} \ , \tag{19}$$

with $e_{ijk} \sim \mathsf{N}(0, \sigma_{e_m}^2)$. In contrast to the corrected model, the naive model absorbs the error variance $\sigma_{e_m}^2$ in the environmental variance σ_E^2, inducing biased estimates of heritability as in equation (4). In addition to the upward bias in $\hat{\sigma}_E^2$, the naive model also leads to an upward bias in $\hat{\sigma}_{PE}^2$, because the repeated measurements within sessions (*i.e.* during constant environmental conditions) induce an increased within-animal correlation that is absorbed by the permanent environmental variance. In all models, priors were inverse-Gammas with variance parameter set to 1 and degree of belief to 0.002, which can be reparameterized with shape and scale parameters equal to 0.001. We did a sensitivity check by using different priors, but the results did not change qualitatively.

Selection. For each simulated pedigree, selection gradients were calculated from the error-free and the error-prone trait, the latter with and without the error correc-

tion. The actual estimate of the selection gradient was obtained by fitting a linear regression model, with the relative fitness trait as response variable and the phenotypic trait and sex as explanatory variables. The same model was fitted using the error-prone phenotype to derive a naive estimate of selection. The error-corrected selection was estimated using a three-layer hierarchical model as given by (11)-(13), again with sex as explanatory variable.

Response to selection. For each simulated pedigree, response to selection was estimated via both R_{BE} and R_{STS}. For the former, the heritabilities estimated from models (17)-(19) were multiplied by the selection differentials, which were the estimated phenotypic covariances $\hat{\sigma}_p(\boldsymbol{z}, \boldsymbol{w})$ for the actual case, and $\hat{\sigma}_p(\boldsymbol{z}^\star, \boldsymbol{w})$ for the naive and corrected cases. To estimate the genetic covariance between the trait and fitness, $\sigma_a(\boldsymbol{z}, \boldsymbol{w})$, we used the bivariate versions of models (17)-(19), exactly as formulated in equations (14) and (15). Parameter-expanded priors were used for the variance components in order to facilitate convergence of parameters that were close to zero, according to Hadfield (2010). The working parameter prior was normally distributed, with a mean of 0 and a variance of 1000. The prior for the error variance was fixed at 0 for the fitness trait, because no error modeling is required in that component.

Body mass of snow voles

The empirical data we use here stem from a snow vole population that has been monitored between 2006 and 2014 in the central eastern Swiss Alps (Bonnet et al., 2017). The genetic pedigree is available for 937 voles, together with measurements on morphological and life history traits. Thanks to the isolated location, it was possible to monitor the whole population and to obtain relatively high recapture probabilities (0.924 ± 0.012 for adults and 0.814 ± 0.030 for juveniles). Details of the study are given in Bonnet et al. (2017).

Our analyses focused on the estimation of quantitative genetic parameters for the animals' body mass (m, in grams). The dataset contains 3266 mass observations from 917 different voles across 9 years. Such measurements are expected to suffer from classical measurement error, as they were taken with a spring scale, which is prone to measurement error under field conditions. In addition, the actual mass of an animal may fluctuate by transient within-day effects (eating, defecating, digestive processes), as well as by unknown pregnancy conditions in females, which cannot reliably be determined in the field. Repeated measurements are available, both recorded within and across different seasons, and the number of capturing sessions per season varied between 2 and 5 per year. Note that such sessions were defined

by purely operational aspects when the data were collected. Implications for error modeling and ways to handle them will be further discussed in the following sections.

Heritability

Bonnet et al. (2017) estimated heritability using an animal model with sex, age, Julian date (JD), squared Julian date and the two-way and three-way interactions among sex, age and Julian date as fixed effects. The inbreeding coefficient was included to avoid bias in the estimation of additive genetic variance (de Boer and Hoeschele, 1993). The breeding value (a_i), the maternal identity ($mother_i$) and the permanent environmental effect explained by the individual identity (id_i), were included as individual-specific random effects. Here we interpret repeats taken within four consecutive trapping nights (with a mean break of three weeks between) as short-term repeats, namely because this was the length of a "trapping session" during the data collection process. It is possible that four days might be undesirably long, and that variability in such an interval thus includes more than purely transient effects, but the data did not allow for a finer time-resolution. Despite this, and to illustrate the importance of the session length, we also repeated the analysis with sessions defined as a calendar month, which is expected to identify a larger (probably too high) proportion of variance captured by $\sigma^2_{e_m}$. The number of 4-day sessions per individual was on average 2.89 (min = 1 , max = 19) with 1.15 (min = 1, max = 3) number of short-term repeats on average, while the there were 2.28 (min = 1 , max = 12) 1-month sessions on average, with 2.16 (min = 1, max = 8) short-term repeats per session.

If the analyst does not discriminate between short-term (within session) and long-term (across sessions) repeated measurements, the naive model is given as

$$m_{ijk} \;=\; \mu + \boldsymbol{X}_{ijk}\boldsymbol{\beta} + a_i + mother_i + id_i + E_{ijk} \;, \tag{20}$$

where m_{ij} is the mass of animal i in session j for repeat k. Apart from the maternal effect used to disentangle the maternal variance σ^2_M, this model is the analogon to equation (18) and is thus prone to underestimate heritability, because the error variance cannot separate $\sigma^2_{e_m}$ from the total variability. To correctly isolate the measurement error variance, the same model expansion as in (19) leads to the corrected model

$$m_{ijk} \;=\; \mu + \boldsymbol{X}_{ijk}\boldsymbol{\beta} + a_i + mother_i + id_i + E_{ij} + e_{ijk} \;,$$

with $E_{ij} \sim \mathsf{N}(0, \sigma^2_E)$ and $e_{ijk} \sim \mathsf{N}(0, \sigma^2_{e_m})$. Under the assumption that the length of a session was defined in an appropriate way, this model yields an unbiased estimate

of h^2, as given in equation (2). In both models, inverse Wishart priors with variance parameter set to 1 and degree of belief to 0.002 were used for all variance components. In the corrected model, the error variance was given an inverse Wishart prior with variance parameter set to 1 and degree of belief to 0.002. Analyses were repeated with an inverse Wishart prior with variance parameter set to 4 (*i.e.* variance among repeats) and degree of belief to 0.002 for a sensitive check on priors and results did not change qualitatively.

Selection

Selection was estimated from the regression of relative fitness on the actual and error-prone body mass. Relative fitness $\boldsymbol{w}$ was defined as the relative lifetime reproductive success (rLRS), calculated as the number of offspring over the lifetime of an individual, divided by the population mean LRS. The naive estimate of the selection gradient was obtained from a linear mixed model, where the phenotypic trait, sex and age were included as fixed effects, plus a cohort-specific random effect. The error-corrected selection was estimated using a three-layer hierarchical model as in (11)-(13), but with the additional random effect for cohort in the regression model. Sex and age were also included as fixed effects in the exposure model, plus breeding values, environmental and permanent environmental effects as random effects. The full hierarchical model used to estimate the corrected selection gradient is reported in Appendix 1. As pointed out by Bonnet et al. (2017), rLRS is not actually a Gaussian trait, which means that the CIs of the estimate for selection (β_z) are incorrect. We therefore also used an over-dispersed Poisson regression model for both the naive and the full corrected hierarchical model with absolute LRS as the response to calculate p-values based on the correct assumptions. Posterior predictive p-values were used for the hierarchical corrected models. The zero-inflated Poisson models are reported in Appendix 1, and the R code for the implementation of all models is reported in Appendix 4.

Response to selection

Response to selection was estimated using both the breeder's equation (1) and the secondary theorem of selection (3), and in each case using naive and error-corrected models. The naive and corrected values of R_{BE} were estimated by substituting either the naive h^2_{biased} or the corrected h^2 into the breeder's equation. On the other hand, R_{STS} was estimated from the bivariate animal model using the same fixed and random effects as those used in the estimation of heritability in equation (20). The residual variance for the fitness trait was set to zero, so that the residual variance in fitness is correctly captured by the permanent environment random effect, and not

by the environmental variance (Bonnet et al., 2017). Please refer to Appendices 3, 4 and 5 for the R code of all examples.

Results

Heritability

Simulations

As predicted by theory, the naive models yielded lower estimates of heritability, and this bias increased with increasing error variance (Figure 2). The corrected models were able to recover estimates of heritability that are much closer to the actual simulated values in all three cases. Importantly, simulations confirm that σ^2_A can be estimated unbiased in the presence of independent error, even when the error variance is large. As expected, however, the permanent environmental variance σ^2_{PE} is upward biased in the naive model. In addition, we see an increase in the environmental component of variance, and the bias grows with the error variance, namely because $\hat{\sigma}^2_E$ then absorbs also the error variance (see equations (18) and (19)). The respective corrected models succeeded in disentangling all these components of variance and in retrieving the initial simulated value of environmental variance, as well as a correct estimate of the error variance itself (see Table 1 and 2 in Appendix 1).

Body mass of snow voles

An overview of all variance and heritability estimates with their respective 95% highest posterior credible intervals (CIs) from the MCMC runs is given in Table 1. The estimates of the additive genetic variance σ^2_A, and the maternal variance σ^2_M are relatively stable when comparing naive and corrected models, while the permanent environmental variance σ^2_{PE} is again overestimated in the naive model. On the other hand, the environmental variance is split into environmental and transient (error) variance when the corrected model is used, so that $\hat{\sigma}^2_E + \hat{\sigma}^2_{e_m}$ in the corrected model corresponds approximately to $\hat{\sigma}^2_E$ from the naive model.

Table 1 also contains the corrected values when sessions are defined as a full calendar month. As expected, the estimated error variance is smaller when a session is defined as a 4-day interval ($\hat{\sigma}^2_{e_m} = 8.13$) than as a full month ($\hat{\sigma}^2_{e_m} = 10.66$), although the latter certainly encompasses also environmental, non-transient variability of the trait. As a consequence, the correction for heritability is stronger when the longer session definition is used. This example is instructive because it underlines the importance of defining the time scale at which short-term repeats are expected to

capture only transient variance, and not actual variability of the phenotypic trait. In the case of the mass of a snow vole, most biologists would probably agree that a 1-month session is unreasonably long, while it is less clear how much of the fluctuations within a 4-day session are transient, and what part of it would be relevant for selection. Within-day repeats might be the most appropriate for the case of mass, at least when within-day variance is regarded as transient, but such repeats were unfortunately not available here in sufficient quantity, namely because the data were not actually collected with the intention to quantify such effects. Thus, because these data are retrospectively used for analyses that the data collection process was not designed for, we acknowledge that these results are only valid to give conceptual insight, while $\hat{\sigma}^2_{e_m}$ at the 4-day interval is potentially still overestimating the actual value of the transient variance.

Selection

Simulations

The simulation study confirmed that selection gradients are underestimated in the presence of error. The downward bias increased with increasing error variances and naive models yielded progressively lower estimates of selection. On the other hand, the corrected models provided unbiased estimates of selection (see Figure 3 and Table 1 and 2 in Appendix 1.) Even though in the case of an error variance equal to 0.5 the corrected values show an increase from the true values, possibly due to sampling error, the estimates from the corrected models are closer to the true values than the naive estimates and are not affected by the presence of the error.

Body mass of snow voles

Table 2 displays the estimates of selection gradients ($\hat{\beta}_z$) obtained with the naive and corrected models. The error-corrected model, as expected, provided a higher estimate of selection than the naive model, with an increase from 0.03 to 0.07 (4-day session) and to 0.23 (1-month session). The definition of sessions also has an impact on the correction, with longer sessions leading to higher estimates $\hat{\sigma}^2_{e_m}$ and therefore a stronger correction for selection $\hat{\beta}_z$. The p-values of the zero-inflated Poisson models confirmed the effect of selection in the naive ($p < 0.001$) and in the corrected models ($p < 0.001$ for both corrections).

Response to selection

Simulations

Estimates for the response to selection estimated with the breeder's equation ($\hat{R}_{\mathrm{BE}}$) suffered from the bias propagated by the estimate of heritability, and the bias disappeared when introducing the corrected estimate of heritability (Figure 4). On the other hand, the error in the phenotype does not have a noticeable effect on the estimates of response to selection estimated by the secondary theorem of selection ($\hat{R}_{\mathrm{STS}}$), which is in line with theoretical expectations.

Body mass of snow voles

Table 3 contains the estimated values for the evolutionary response to selection for the different combinations of models (naive, corrected) and approaches (breeder's equation, secondary theorem of selection). When using the 4-day session definition, the error-corrected model estimated a stronger response to selection than the naive model for the breeder's equation, with a change from $\hat{R}_{\mathrm{BE}} = 0.10$ to 0.16, which is in line with theory. However, in contrast to the theoretical expectation that predicts that R_{STS} is not affected by the error, the respective estimate shows a similar error correction effect (from -0.33 to -0.37). Only the 1-month definition of the session illustrates that $\hat{R}_{\mathrm{BE}}$ is clearly more bias-corrected (from 0.10 to 0.20) than $\hat{R}_{\mathrm{STS}}$ (still from -0.33 to -0.37) – although this is probably an over-correction of R_{BE} under the premise that a full month cannot be regarded as a short-term interval for the mass trait. Still, the example illustrates that the breeder's equation is generally more prone than the secondary theorem of selection to underestimate selection mechanisms in real study systems when measurement error in the phenotype is present (Table 1). On the other hand – and probably more importantly – the results confirm that estimates for response to selection may differ dramatically between the breeder's equation and the secondary theorem of selection approach. As already noticed by Bonnet et al. (2017), the predicted evolutionary response derived from the breeder's equation may even point in the opposite direction of the estimate derived by the secondary theorem of selection (*e.g.* naive estimates $\hat{R}_{\mathrm{BE}} = 0.10$ vs. $\hat{R}_{\mathrm{STS}} = -0.33$, with non-overlapping CIs).

Discussion

We have addressed the problem of mismeasured continuous phenotypic traits, when these are included in quantitative genetics analyses, and suggested modeling strategies to obtain bias-corrected estimates of heritability, selection gradients and the response to selection. Such strategies rely on the distinction between variability

from (stable) environmental effects that are part of the phenotypic variability, and (transient) error effects. We argued that ignoring such a distinction may not only lead to an *under*estimation of the actual heritability (as pointed out by Ge et al., 2017), but also to bias in the estimates of selection gradients and the response to selection. The concept of "transient effects" can be understood as a generalization of measurement error, and it may include transient fluctuations in a broader sense, *e.g.* periodic within-day variations. Repeated measurements help to isolate the variance from such transient fluctuations, which in turn allows to correct for bias.

On the other hand, the use of repeats to estimate variance parameters is nothing new to quantitative geneticists, see *e.g.* Wilson et al. (2010), but these were typically employed to separate permanent environmental effects from the additive genetic variance, which prevents h^2 from being *over*estimated. The fact that repeated measurements are used to correct for opposite biases in heritability estimates makes it apparent that the information content in what is termed "repeated measurements" in both cases is indeed very different. The crucial aspect, which we attempted to clarify in this article, is that it matters at which temporal distance the repeats were taken, and that the relevance of this distance depends on the kind of trait under study. For a plastic phenotypic trait, repeats taken on the same individual at different life stages ("long-term" repeats, *e.g.* across measurement sessions) can be used to separate the animal-specific permanent environmental effect from both genetic and environmental variances. On the other hand, repeats taken in temporal vicinity ("short-term" repeats, *e.g.* within a measurement session) may disentangle pure measurement error and other transient fluctuations from environmental effects, if it can be assumed that the environmental component does not change in the short time interval between the measurements. Only in the presence of *both* types of repeats, that is, across different (and relevant) time scales, is it practically feasible to separate all four variance components. The respective mixed model for the trait value, typically the animal model, then needs extension to three levels of measurement hierarchy: the animal (i), the session (j within i) and the repeat (k within j within i), as suggested in equation (9). However, it is not possible to give general guidelines on how to discriminate between long-term and short-term repeats, because this must be answered by the investigator, using biological knowledge of the system and the trait. In general, if one is unsure if repeats capture transient or persistent effects, one might ask if averaging over repeats, as proposed by certain authors (Carbonaro et al., 2009; Zheng et al., 2016), is expected to result in better estimates of quantitative genetics measures or not, because such averaging methods have exactly the purpose to reduce bias that emerges due to transient fluctuations. Note that this does not mean that we advocate to average over repeated measurements in most cases, because such a practice can only alleviate bias (by reducing the

error variance in the mean), but not eliminate it completely. We suggest instead to define biologically relevant time-scales for repeated measurements, and to explicitly model all relevant variance components.

Our method approaches the problem by assuming a dichotomous distinction between short-term and long-term repeats. Another way to look at within-animal repeated measurements is by recalling that these are usually correlated, even when taken across long time spans. On the other hand, when repeats are taken at close temporal scales, the similarity of the environmental component induces an even higher correlation, at least when the environment is expected to change only slowly (*e.g.* due to seasonal effects). A more sophisticated model could take into account that environmental conditions, and thus the environmental component E_i in the model, change continuously, such that the correlations of repeated measurements declines over time. This kind of model might be beneficial if repeats were not taken in clearly defined measurement sessions, because it would not rely on the dichotomy between short-/long-term repeats. Of course, introducing such a temporal correlation term introduces another level of complexity, and thus entails other challenges.

In any case, given the importance of repeated measurements and the information they capture, it becomes apparent that a carefully evaluated study design is critical. Still, it may sometimes not be possible to take a sufficient number of measurements on the same individual. It is thus important to notice that even in the absence of short-term repeats, the inclusion of the respective random effect may be feasible, provided that knowledge about the error variance is available, *e.g.* from previous studies, from a subset of the data, or from other "expert" knowledge. The Bayesian framework is ideal in this regard, because it is straightforward to include random effects with a very strong (or fixed) prior on the respective variance component.

We have seen that for the same reason that heritability is systematically underestimated in the presence of independent phenotypic error, selection gradients and the response to selection (when estimated by the breeder's equation, $\hat{R}_{\mathrm{BE}}$) are underestimated as well, while the Robertson-Price identity (*i.e.* $\hat{R}_{\mathrm{STS}}$) does not suffer from a bias. Other reasons to prefer the latter when estimating response to selection were given in Morrissey et al. (2010) and Morrissey et al. (2012), while Engen and Seather (2013) and Engen et al. (2014) also suggest a decomposition of different effects in the estimation of such parameter based on the distinction of transient effects. Nevertheless, the Robertson-Price identity does not model selection explicitly and says little about selective processes. The Robertson-Price equation can thus be used to check the consistency of predictions made from the breeder's equation, but the breeder's equation remains necessary to test hypothesis about the causal nature of selection (Morrissey et al., 2012; Bonnet et al., 2017). Besides, a critical assumption of our models was that the transient (error) components are independent of the true

covariate, but also independent of other covariates or the (fitness) trait. While the small changes in $\hat{R}_{\mathrm{STS}}$ that were observed in the snow voles application could be due to pure stochasticity, a non-neglectable alternative is that the measurement error in the data is not independent. In fact, it is quite likely that the error in mass is correlated with fitness, because some of the error may be due to undetected pregnancy in females. For pregnant animals, values for body mass are probably overestimated, and we expect the respective (positive) deviation from the actual value to be correlated with fitness, because *given* that an individual is pregnant at a specific time point, it has a higher expected number offspring over its entire lifespan.

The discussion in this article has mainly revolved around traits that display a relatively high level of phenotypic plasticity. However, the considerations fall somewhat short for static traits, for example the sceletal size of a fully grown animal, where repeated measures, even years apart, will yield meaningful estimates of measurement error (except maybe at very high age). This thought-experiment illustrates that traits without plasticity facilitate the isolation of measurement error, but it then becomes impossible to separate permanent environmental effects from genetic effects by using within-individual repeats, exactly because repeats will always capture error and not environmental changes. For such traits, only the inclusion of additional effects, such as maternal or other shared-environmental effects between animals, allows to disentangle σ^2_{PE} from σ^2_A, and they may still prove insufficient, *e.g.* when the same environment is shared between relatives and temporal, brood or prenatal effects might be required to properly separate the different variance components (Kruuk and Hadfield, 2007; Hadfield et al., 2013). In general, the importance of separating the right variances, and in particular their inclusion in the denominator of heritability estimates, has been intensely discussed in the past, but the focus was mainly on the proper handling of variances captured by fixed effects (Wilson, 2008; de Villemereuil et al., 2017). Note, however, that, we have not discussed non-additive genetic effects that occur in the presence of interaction between genes, such as dominance or epistatis. These are likely to affect the estimates of heritability and also need to be taken into account when estimating variance components and heritability of such traits (Meffert et al., 2002; Palucci et al., 2007).

Furthermore, the methods presented in this paper have been developed and implemented on continuous phenotypic traits. Of course, binary, categorical or count traits might as well suffer from measurement error, which is then denoted as misclassification error (Copas, 1988; Magder and Hughes, 1997; Küchenhoff et al., 2006), or as miscounting error (Muff et al., 2017). Models for non-Gaussian traits are usually formulated in a generalized linear model framework (Nakagawa and Schielzeth, 2010; de Villemereuil et al., 2016) and require the use of a link function (*e.g.* the logistic or log link). The phenotypic variance can then be decomposed on the latent scale,

and it might be possible to include an error term e_{ijk} as in equation (9). However, we believe that misclassification and miscounting error should be addressed with extended modeling strategies, such as hierarchical models with an explicit level for the error process.

We hope that the considerations and methods provided here serve as a useful starting point to build on when estimating quantitative genetics parameters in the presence of mismeasured phenotypic traits or transient effects. Importantly, all approaches proposed here are relatively straightforward to understand and implement, but further generalizations are possible and will hopefully follow in the future.

Supporting information:

Appendix 1: Supplementary text and figures (pdf)

Appendix 2: R script for the simulation and analysis of pedigree data

Appendix 3: R script for heritability in snow voles

Appendix 4: R script for selection in snow voles

Appendix 5: R script for response to selection in snow voles.

Acknowledgements:

This work would not have been possible without generous funding from the Faculty of Science of the University of Zurich and from the grant Forschungskredit of the University of Zurich, grant no. FK-16-097.

Data archive location:

The snow voles data used here have previously been archived on https://doi.org/10.1371/journal.pbio.1002592.

References

Bonnet, T., P. Wandeler, G. Camenisch, and E. Postma (2017). Bigger is fitter? Quantitative genetic decomposition of selection reveals an adaptive evolution decline of body mass in a wild rodent population. *PLOS Biology 15*, e1002592.

Carbonaro, F., T. Andrew, D. Mackey, T. Young, T. Spector, and C. Hammond (2009). Repeated measures of intraocular pressure result in higher heritability and greater power in genetic linkage studies. *Investigative Ophthalmology and Visual Science 50*, 5115–5119.

Carroll, R., D. Ruppert, L. Stefanski, and C. Crainiceanu (2006). *Measurement error in nonlinear models, a modern perspective*. Boca Raton: Chapman and Hall.

Charmantier, A., D. Garant, and L. E. B. Kruuk (2014). *Quantitative Genetics in the Wild*. Oxford: Oxford University Press.

Charmantier, A. and D. Reale (2005). How do misassigned paternities affect the estimation of heritability in the wild? *Molecular Ecology 14*, 2839–2850.

Copas, J. B. (1988). Binary regression models for contaminated data (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 50*, 225–265.

Coster, A. (2013). *pedigree: Pedigree functions*. R package version 1.4.

de Boer, I. and I. Hoeschele (1993). Genetic evaluation methods for populations with dominance and inbreeding. *Theoretical Applied Genetics 86*, 245–258.

de Villemereuil, P., M. B. Morrissey, S. Nakagawa, and H. Schielzeth (2017). Fixed effect variance and the estimation of the heritability: Issues and solutions. *bioRxiv*. http://www.biorxiv.org/content/early/2017/07/04/159210.

de Villemereuil, P., H. Schielzeth, S. Nakagawa, and M. Morrissey (2016). General methods for evolutionary quantitative genetic inference from generalized mixed models. *Genetics 204*, 1281–1294.

Dohm, M. (2002). Repeatability estimates do not always set an upper limit to heritability. *Functional Ecology 16*, 273–280.

Engen, S., T. Kvalnes, and B. Seather (2014). Estimating phenotypic selection in age-structured populations by removing transient fluctuatuions. *Evolution 68*, 2509–2523.

Engen, S. and B. Seather (2013). Evolution in fluctuating eenvironment: decomposition into additive components of the Robertson-Price equation. *Evolution 68*, 854–865.

Falconer, D. and T. Mackay (1996). *Introduction to Quantitative Genetics*. Burnt Mill, Harlow, Essex, England: Pearson.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford, UK: Oxford University Press.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

Ge, T., A. Holmes, R. Buckner, J. Smoller, and M. Sabuncu (2017). Heritability analysis with repeat measurements and its application to resting-state functional connectivity. *PNAS 114*, 5521–5526.

Griffith, S., I. Owens, and K. A. Thuman (2002). Extrapair paternity in birds: a review of interspecific variation and adaptive function. *Molecular Ecology 11*, 2195–2212.

Hadfield, J. (2008). Estimating evolution parameters when viability selection is operating. *Proceedings of the Royal Society of London B: Biological Sciences, The Royal Society 275*, 723–734.

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

Hadfield, J., E. Heap, F. Bayer, E. Mittell, and N. Crouch (2013). Disentangling genetic and prenatal sources of familial resemblance across ontogeny in a wild passerine. *Evolution 67*, 2701–2713.

Henderson, C. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics 32*, 69–83.

Hill, W. G. (2014). Applications of population genetics to animal breeding, from wright, fisher and lush to genomic prediction. *Genetics 196*, 1–16.

Keller, L. and A. Van Noordwijk (1993). A method to isolate environmental effects on nestling growth, illustrated with examples from the Great Tit (Parsus major). *Functional Ecology 7*, 493–502.

Keller, L. F., P. R. Grant, B. R. Grant, and K. Petren (2001). Heritability of morphological traits in Darwin's Finches: misidentified paternity and maternal effects. *Heredity 87*, 325–336.

Kruuk, L. (2004). Estimating genetic parameters in natural populations using the 'animal model'. *Philosophical Transactions of the Royal Society B: Biological Sciences 359*, 873–890.

Kruuk, L. and J. Hadfield (2007). How to separate genetic and environmental causes of similarity between relatives. *Journal of Evolutionary Biology 20*, 1890–1903.

Küchenhoff, H., S. Mwalili, and E. Lesaffre (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics 62*, 85–96.

Lande, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution 33*, 402–416.

Lande, R. and S. J. Arnold (1983). The measurement of selection on correlated characters. *Evolution 37*, 1210–1226.

Logan, C., L. Kruuk, R. Stanley, A. Thompson, and T. Clutton-Brock (2016). Endocranial volume is heritable and is associated with longevity and fitness in a wild mammal. *Royal Society Open Science 3*, 160622.

Lush, J. (1937). *Animal breeding plans.* Ames, Iowa: Iowa State College Press.

Lynch, M. and B. Walsh (1998). *Genetics and Analysis of Quantitative Traits.* Sunderland, MA: Sinauer Associates.

Macgregor, S., B. Cornes, N. Martin, and P. M. Visscher (2006). Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Human Genetics 120*, 571–580.

Magder, L. S. and J. P. Hughes (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology 146*, 195–203.

Meffert, L., S. Hicks, and J. Regan (2002). Nonadditive genetic effects in animal behavior. *The American Naturalist 160 Suppl 6*, S198–S213.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics 157*, 1819–1829.

Mitchell-Olds, T. and R. Shaw (1987). Regression analysis of natural selection: statistical inference and biological interpretation. *Evolution 41*, 1149–1161.

Møller, A. and M. D. Jennions (2002). How much variance can be explained by ecologists and evolutionary biologists? *Oecologia 132*(4), 492–500.

Morrissey, M., D. Parker, P. Korsten, J. Pemberton, L. Kruuk, and A. Wilson (2012). The prediction of adaptive evolution: empirical application of the secondary theorem of selection and comparison to the breeder's equation. *Evolution 66*, 2399–2410.

Morrissey, M. B., L. Kruuk, and A. Wilson (2010). The danger of applying the breeder's equation in observational studies of natural populations. *Journal of Evolutionary Biology 23*, 2277–2288.

Muff, S., M. A. Puhan, and L. Held (2017). Bias away from the Null due to miscounted outcomes? a case study on the TORCH trial. *Statistical Methods in Medical Research*. In press.

Muff, S., A. Riebler, L. Held, H. Rue, and P. Saner (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society- Applied Statistics 64*, 231–252.

Nakagawa, S. and H. Schielzeth (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews of the Cambridge Philosophical Society 85*, 935–956.

Notter, D., J. Burke, J. Miller, and J. Morgan (2017). Factors affecting fecal egg counts in periparturient Katadhin ewes and their lambs. *Journal of Animal Science 95*, 103–112.

Palstra, F. and D. Fraser (2012). Effective/census population size ratio estimation: a compendium and appraisal. *Ecology and Evolution 2*, 2357–2365.

Palucci, V., L. R. Schaeffer, F. Miglior, and V. Osborne (2007). Non-additive genetic effects for fertility traits in Canadian Holstein cattle. *Genetics Selection Evolution 39*, 181–193.

Peek, M. S., A. J. Leffler, S. D. Flint, and R. J. Ryel (2003). How much variance is explained by ecologists? Additional perspectives. *Oecologia 137*(2), 161–170.

Ponzi, E., L. Keller, and S. Muff (2017). Pedigree reconstruction: the simulation extrapolation technique meets quantitative genetics. Technical report, University of Zürich.

Price, G. R. (1970). Selection and covariance. *Nature 227*, 520–521.

Price, T. D. and P. T. Boag (1987). Selection in natural populations of birds. In F. Cooke, , and P. Buckley (Eds.), *Avian Genetics*, pp. 257 – 287. Academic Press.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richardson, S. and W. Gilks (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine 12*, 1703–1722.

Robertson, A. (1966). A mathematical model of the culling process in dairy cattle. *Animal Science 8*, 95–108.

Roff, D. A. (2007). A centennial celebration for quantitative genetics. *Evolution 61*, 1017–1032.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society Series B (Statistical Methodology) 71*, 319–392.

Senneke, S., M. MacNeil, and L. Van Vleck (2004). Effects of sire misidentification on estimates of genetic parameters for birth and weaning weights in Hereford cattle. *Journal of Animal Science 82*, 2307–2312.

Steinsland, I., C. Larsen, A. Roulin, and H. Jensen (2014). Quantitative genetic modeling and inference in the presence of nonignorable missing data. *Evolution 68*, 1735–1747.

Stephens, D. and P. Dellaportas (1992). Bayesian analysis of generalised linear models with covariate measurement error. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 4*. Oxford Univ Press.

van der Sluis, S., M. Verhage, D. Posthuma, and C. Dolan (2010). Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PLOS One 5*, e13929.

Wilson, A., D. Réale, M. Clements, M. Morrissey, E. Postma, C. Walling, L. Kruuk, and D. Nussey (2010). An ecologist's guide to the animal model. *Journal of Animal Ecology 79*, 13–26.

Wilson, A. J. (2008). Why h^2 does not always equal VA/VP? *Journal of Evolutionary Biology 21*, 647–650.

Wolak, M. E. and J. M. Reid (2017). Accounting for genetic differences among unknown parents in microevolutionary studies: how to include genetic groups in quantitative genetic animal models. *Journal of Animal Ecology 86*, 7–20.

Zheng, Y., R. Plomin, and S. von Stumm (2016). Heritability of intraindividual mean and variability of positive and negative affect: genetic analysis of daily affect ratings over a month. *Psychological Science 27*, 1611–1619.

Figures

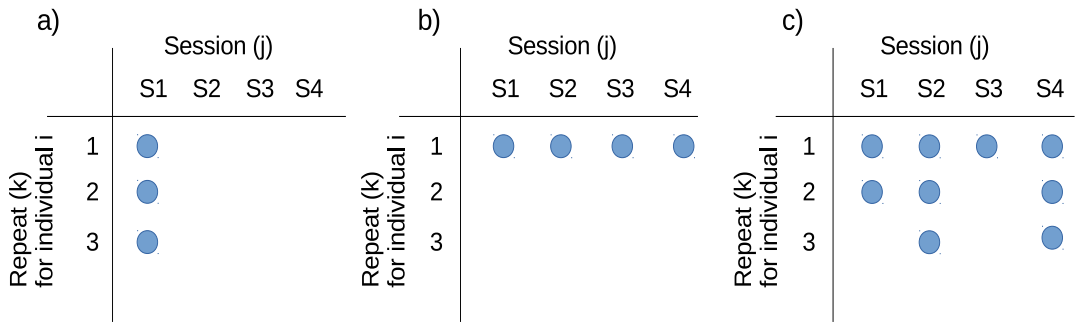

Figure 1: Schematic representation of three study designs where repeated measurements are taken on one individual in a) a single measurement session, b) as single measurements across sessions, or c) as multiple measurements across multiple sessions. Case a) allows to disentangle transient (measurement error) variance $\sigma^2_{e_m}$ from σ^2_E, and case b) allows to separate permanent environmental effects σ^2_{PE} from genetic components σ^2_A. All four variance components can be separated in case c).

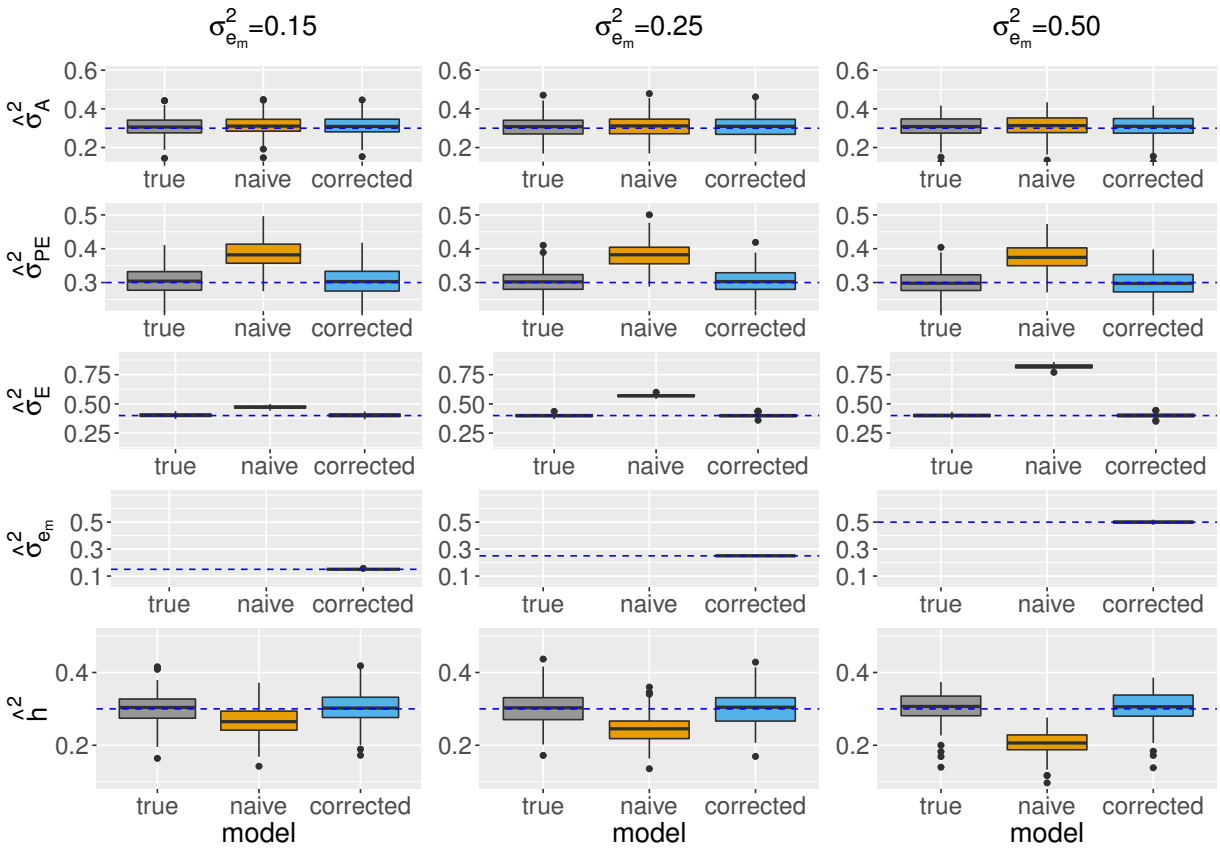

Figure 2: Distribution of the posterior means of the estimated variance components and heritability in the 100 simulated pedigrees. Results are shown for three different values of the error variance. Heritability and variance components are estimated for the true data without measurement error and compared to the data with error by using a naive and a corrected model. The error variance is estimated in the corrected models and compared to its simulated value in each case.

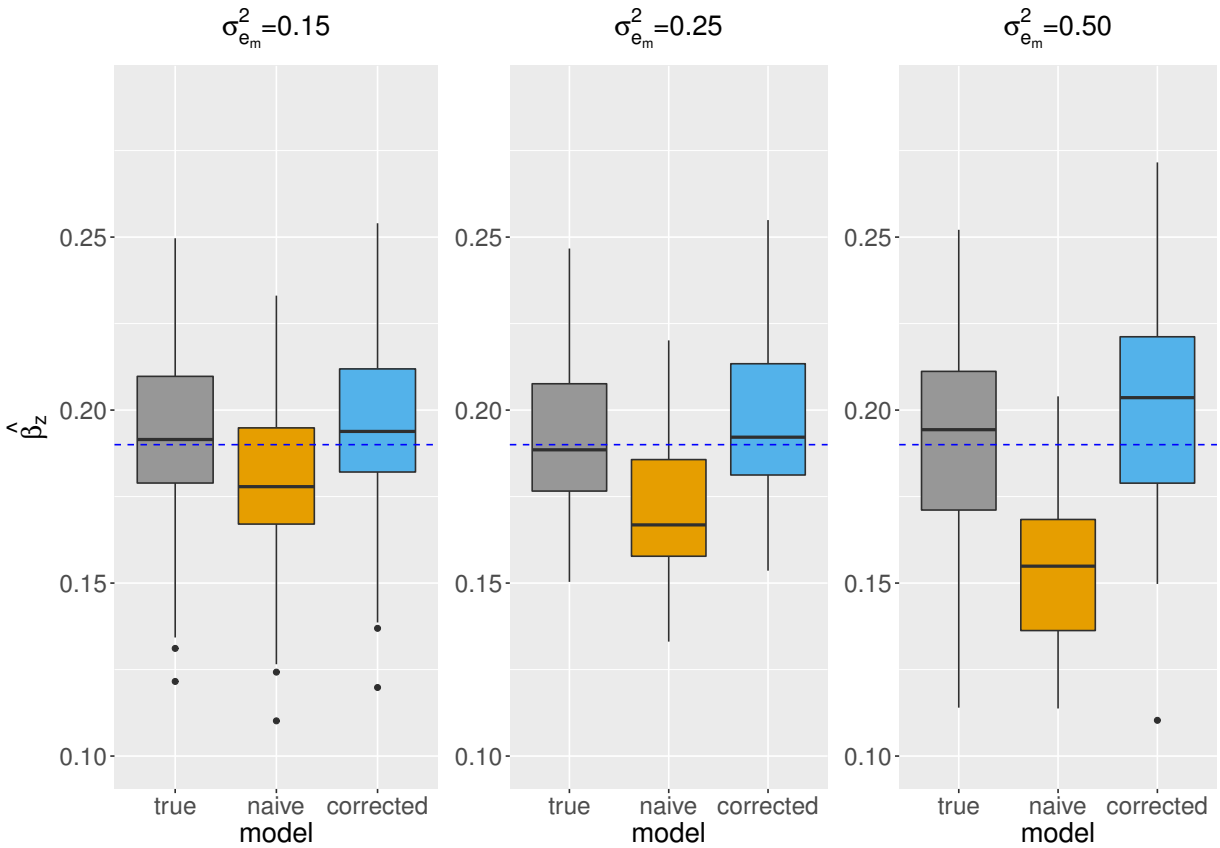

Figure 3: Distribution of the posterior means from the 100 simulated pedigrees for selection. Results are shown for three different values of the error variance. Selection is estimated for the true data without measurement error and compared to the data with error by using first with a naive and then a corrected model.

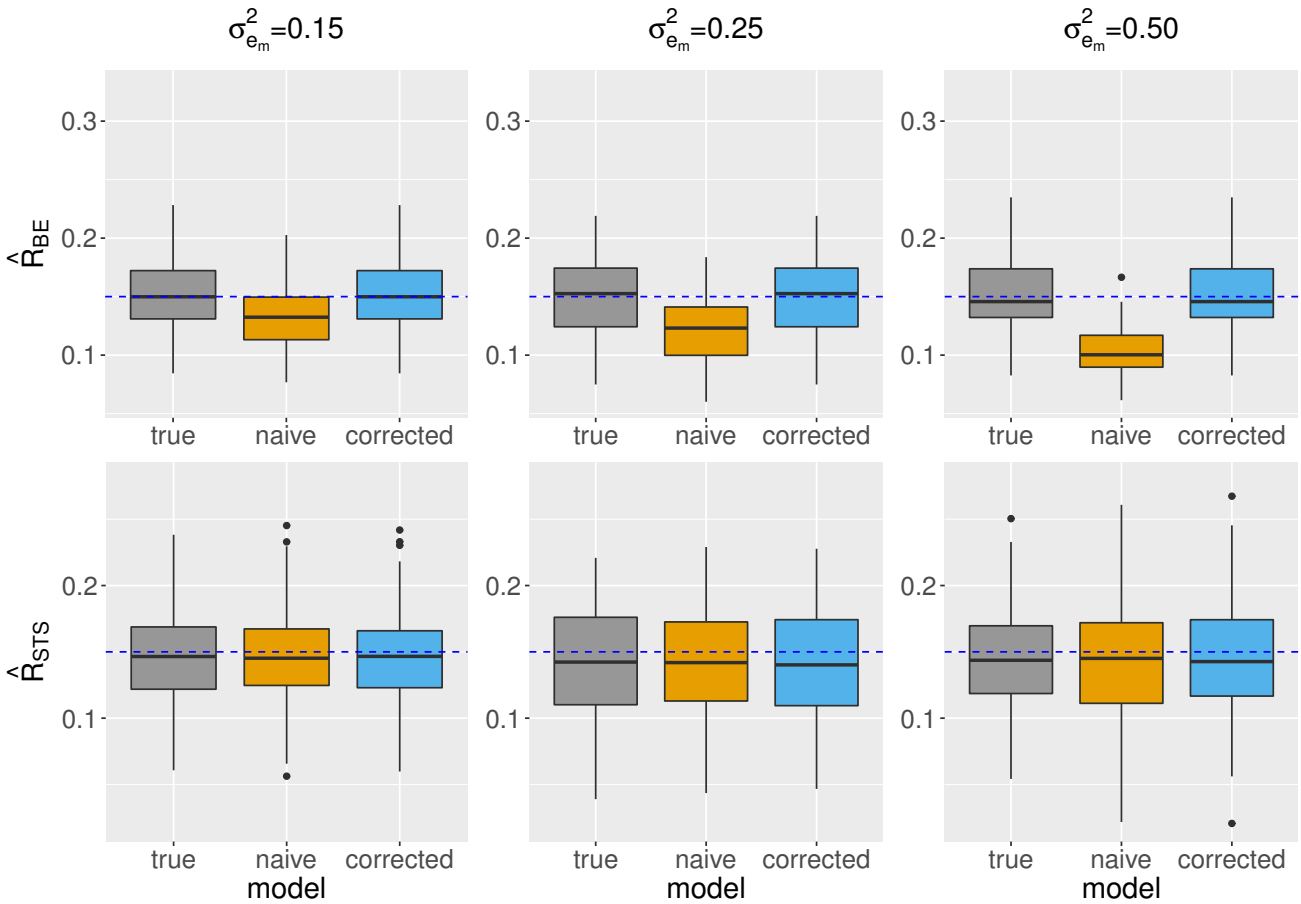

Figure 4: Distribution of the posterior means from the 100 simulated pedigrees for response to selection. Results are shown for three different values of the error variance. Response to selection is estimated for the true data without measurement error and compared to the data with error by using first with a naive and then a corrected model. Estimates are obtained with both the breeder's equation and the secondary theorem of selection.

Tables

	$\hat{h}^2$	$\hat{\sigma}^2_A$	$\hat{\sigma}^2_{PE}$	$\hat{\sigma}^2_M$	$\hat{\sigma}^2_E$	$\hat{\sigma}^2_{e_m}$
naive	0.15	3.50	6.46	1.63	12.46	-
	$[0.06, 0.24]$	$[1.48, 5.94]$	$[4.46, 8.50]$	$[0.67, 2.81]$	$[11.67, 13.13]$	
corrected (4-day session)	0.22	3.45	5.93	1.67	4.73	8.13
	$[0.09, 0.36]$	$[1.36, 5.97]$	$[3.91, 8.13]$	$[0.59, 2.54]$	$[3.76, 5.79]$	$[7.14, 8.90]$
corrected (1-month session)	0.28	3.73	5.45	1.63	2.28	10.66
	$[0.16, 0.44]$	$[1.86, 5.66]$	$[4.02, 7.61]$	$[0.91, 2.64]$	$[1.51, 3.07]$	$[10.13, 11.39]$

Table 1: Estimates of variance components and heritability of body mass in snow voles given by the naive and the corrected models.

	$\hat{\beta}_z$	p-value
naive	0.03	< 0.001
corrected (4-day session)	0.07	< 0.001
corrected (1-month session)	0.23	< 0.001

Table 2: Estimates of selection gradients for body mass in snow voles, derived from the naive and the corrected models. P-values are reported from the zero-inflated Poisson models and calculated as posterior predictive p-values for the hierarchical corrected models.

	$\hat{R}_{\text{STS}}$	95% CI	$\hat{R}_{\text{BE}}$	95% CI
naive	-0.33	$[-0.65, -0.05]$	0.10	[0.04, 0.17]
corrected (4-day session)	-0.37	$[-0.67, -0.06]$	0.16	[0.06, 0.26]
corrected (1-month session)	-0.37	$[-0.69, -0.08]$	0.20	[0.11, 0.32]

Table 3: Response to selection estimated with the breeder's equation ($\hat{R}_{\text{BE}}$) and with the secondary theorem of selection ($\hat{R}_{\text{STS}}$), for body mass in snow voles, as derived from the naive and the corrected models.