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 and Stefanie Muff1,2,5

1Department of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
2Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zürich, Hirschengraben 84, 8001 Zürich, Switzerland
3Zoological Museum, University of Zürich, Karl-Schmid-Strasse 4, 8006 Zürich, Switzerland
4Division of Ecology and Evolution, Research School of Biology, The Australian National University, Acton, Canberra, ACT 2601, Australia
5E-mail: stefanie.muff@uzh.ch

Received January 12, 2018
Accepted July 12, 2018

Quantitative genetic analyses require extensive measurements of phenotypic traits, a task that is often not trivial, especially in wild populations. On top of instrumental measurement error, some traits may undergo transient (i.e., nonpersistent) fluctuations that are biologically irrelevant for selection processes. These two sources of variability, which we denote here as measurement error in a broad sense, are possible causes for bias in the estimation of quantitative genetic parameters. We illustrate how in a continuous trait transient effects with a classical measurement error structure may bias estimates of heritability, selection gradients, and the predicted response to selection. We propose strategies to obtain unbiased estimates with the help of repeated measurements taken at an appropriate temporal scale. However, the fact that in quantitative genetic analyses repeated measurements are also used to isolate permanent environmental instead of transient effects requires that the information content of repeated measurements is carefully assessed. To this end, we propose to distinguish “short-term” from “long-term” repeats, where the former capture transient variability and the latter help isolate permanent effects. We show how the inclusion of the corresponding variance components in quantitative genetic models yields unbiased estimates of all quantities of interest, and we illustrate the application of the method to data from a Swiss snow vole population.

KEY WORDS: Animal model, Breeder’s equation, error variance, permanent environmental effects, quantitative genetics, Robertson–Price identity.

Quantitative genetic methods have become increasingly popular for the study of natural populations in the last decades, and they now provide powerful tools to investigate the inheritance of characters, and 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 a residual component $R$, so that $P = A + R$. For simplicity, nonadditive genetic effects, such as dominance and epistatic effects, are ignored throughout this article, thus the residual component can be thought of as the sum of all environmental effects. This basic model can be extended in various ways (Falconer and Mackay 1996; Lynch and Walsh 1998), with one of the most common
being \( P = A + PE + R \), where \( PE \) captures dependent effects, the so-called permanent environmental effects, while \( R \) captures the residual, independent variance that remains unexplained. Permanent environmental effects are stable differences among individuals above and beyond the permanent differences due to additive genetic effects. In repeated measurements of an individual, these effects create within-individual covariation. To prevent inflated estimates of additive genetic variance, these effects must therefore be modeled and estimated (Lynch and Walsh 1998; Kruuk 2004; Wilson et al. 2010).

This quantitative genetic decomposition of phenotypes is not possible at the individual level in nonclonal organisms, but under the crucial assumption of independence of genetic, permanent environmental, and residual effects, the phenotypic variance at the population level can be decomposed into the respective variance components as \( \sigma_p^2 = \sigma_A^2 + \sigma_{P,E}^2 + \sigma_R^2 \). These variance components can then be used to understand and predict evolutionary change of phenotypic traits. For example, the additive genetic variance \( (\sigma_A^2) \) can be used to predict the response to selection using the Breeder’s equation. It predicts the response to selection \( R_{BE} \) of a trait \( z \) (bold face notation denotes vectors) from the product of the heritability \( (h^2) \) of the trait and the strength of selection \( (S) \) as

\[
R_{BE} = h^2 \cdot S \tag{1}
\]

(Lush 1937; Falconer and Mackay 1996), where \( h^2 \) is the proportion of additive genetic to total phenotypic variance

\[
h^2 = \frac{\sigma_A^2}{\sigma_p^2} \tag{2}
\]

and \( S \) is the selection differential, defined as the mean phenotypic difference between selected individuals and the population mean or, equivalently, the phenotypic covariance \( \sigma_p(z, w) \) between the trait \((z)\) and relative fitness \((w)\). Besides the Breeder’s equation, evolution can be predicted using the secondary theorem of selection, according to which evolutionary change is equal to the additive genetic covariance of a trait with relative fitness, that is,

\[
R_{STS} = \sigma_a(z, w) \tag{3}
\]

(Robertson 1966; Price 1970). Morrissey et al. (2010, 2012) discuss the differences between the Breeder’s equation and the secondary theorem of selection in detail. A major difference is that in contrast to \( R_{BE} \), \( R_{STS} \) only estimates the population’s evolutionary trajectory, but does not measure the role of selection in shaping this evolutionary change.

One measure of the role of selection is the selection gradient, which quantifies the strength of natural selection on a trait. For a normally distributed trait \( z \), it is given as the slope \( \beta_z \) of the linear regression of relative fitness on a phenotypic trait (Lande and Arnold 1983), that is,

\[
\beta_z = \frac{\sigma_p(z, w)}{\sigma_p^2(z)} \tag{4}
\]

where \( \sigma_p^2(z) \) denotes the phenotypic variance of the trait, for which we only write \( \sigma_p^2 \) when there is no ambiguity about what trait the phenotypic variance refers to.

The reliable estimation of the parameters of interest \((h^2, \sigma_p(z, w), \sigma_a(z, w), \text{and } \beta_z)\), and the successful prediction of evolution as \( R_{BE} \) or \( R_{STS} \), require large amounts of data, often collected across multiple generations and with known relationships among individuals in the dataset. For many phenotypic traits of interest, data collection is not trivial, and multiple sources of error, such as phenotypic measurement error, pedigree errors (wrong relationships among individuals), or nonrandomly missing data may affect the parameter estimates. Several studies have discussed and addressed pedigree errors (e.g., Keller et al. 2001; Griffith et al. 2002; Senneke et al. 2004; Charmantier and Reale 2005; Hadfield 2008) and problems arising from missing data (e.g., Steinsland et al. 2014; 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., Hoffmann 2000; 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 \( z \), an error-prone version

\[
z^* = z + e, \quad e \sim N(0, \sigma_{e^2} I) \tag{5}
\]

is measured, where \( e \) denotes an error term with independent correlation structure \( I \) and error variance \( \sigma_{e^2} \) (see Lynch and Walsh 1998, p. 121). As a consequence, the observed phenotypic variance of the measured values is \( \sigma_p^2(z^*) = \sigma_p^2(z) + \sigma_{e^2} \), and thus larger than the actual phenotypic variance. The error variance \( \sigma_{e^2} \) thus must be disentangled from \( \sigma_p^2(z) \) to obtain unbiased estimates of quantitative genetic parameters. However, most existing methods for continuous trait analyses that acknowledged measurement error have modeled it as part of the residual 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, \( P = A + PE + R \), measurement error is absorbed in \( R \), thus \( \sigma_R^2 \) is absorbed by \( \sigma_R^2 \). This practice effectively downwardly biases measures that are proportions of the phenotypic variance, in particular \( h^2 \) and \( \beta_z \).
To see why, let us denote the biased measures as $h^2_e$ and $\beta^*_z$. The biased version of heritability is then given as

$$h^2_e = \frac{\sigma^2_A}{\sigma^2_P + \sigma^2_e} \leq \frac{\lambda^2}{\lambda^2 + \sigma^2_e}$$

because under the assumption taken here that measurement error is independent of the actual trait value, measurement error is also independent of additive genetic differences and therefore leaves the estimate of the additive genetic variance $\sigma^2_A$ unaffected. This was already pointed out, for example, by Lynch and Walsh (1998, p. 121) or Ge et al. (2017). Equation (6) directly illustrates that $h^2_e$ is attenuated by a factor $\lambda = \sigma^2_p/(\sigma^2_p + \sigma^2_e)$, denoted as reliability ratio (e.g., Carroll et al. 2006). Using the same argument, one can show that $\beta^*_z = \lambda \beta_z$, but also $R^*_BE = \lambda R_{BE}$, as will become clear later.

To obtain unbiased estimates of $h^2$, $\beta_z$, or any other quantity that depends on unbiased estimates of $\sigma^2_p$, it is thus necessary to disentangle $\sigma^2_e$ from the actual phenotypic variance $\sigma^2_P$, and particularly from its residual component $\sigma^2_e$. Importantly, however, purely mechanistic measurement imprecision is often not the only source of variation that may be considered irrelevant for the mechanisms of inheritance and selection in the system under study. Here, we therefore follow Ge et al. (2017) and use the term “transient effects” for the sum of measurement errors plus any biological short-term changes of the phenotype itself that are not considered relevant for the selection process, briefly denoted as “irrelevant fluctuations” of the actual trait.

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 of the magnitude of these effects see Keller and Van Noordwijk 1993). Such short-term fluctuations might not be of interest for the study of evolutionary dynamics, if the fluctuations do not contribute to the selection process in a given population. Under the assumption that these fluctuations are additive and independent among each other and of the actual trait value, they are mathematically indistinguishable from pure measurement error. In the remainder of the article, we therefore do not introduce a separate notation to discriminate between (mechanistic) measurement error and biological short-term fluctuations, but treat them as a single component (e) with a total “error” variance $\sigma^2_e$. Consequently, we may sometimes refer to “measurement error” when in fact we mean transient effects as the sum of measurement error and transient fluctuations.

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 and irrelevant fluctuations of a trait, building on the work by Ge et al. (2017). We start by discussing 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 life span that are relevant for selection. We point out that repeated measures taken over different time intervals can help separate transient effects from more stable (permanent) environmental and genetic effects, and that based on such a variance decomposition it is possible to formulate models that yield unbiased estimates of heritability, selection, and the response to selection. We illustrate these approaches with empirical quantitative trait analyses of body mass measurements taken in a population of snow voles in the Swiss alps (Bonnet et al. 2017).

**Theory**

**SHORT- AND LONG-TERM REPEATED MEASUREMENTS**

Table 1 gives an overview of how the different parameters considered here are (or are not) affected by the presence of measurement error. To retrieve unbiased estimates of all quantities given in Table 1, we must be able to appropriately model and estimate the measurement error variance $\sigma^2_e$, which can be achieved with repeated measurements. These repeated measurements must be taken in close temporal vicinity, that is, on a time scale where the focal trait is not actually undergoing any phenotypic changes that are considered relevant for selection. We introduce the notion of a *measurement session* for such short-term time intervals. In other words, a measurement session can be defined as a sufficiently short period of time during which the investigator is willing to assume that the residual component is constant. On the other hand, measurements are often repeated across much longer periods of time, such as months, seasons, or years, during which phenotypic change is not expected to be solely due to transient effects, and the resulting trait variation is often relevant for selection. Thus, long-term repeats, taken across different measurement sessions, help separating permanent environmental effects from residual components (e.g., Wilson et al. 2010).

The distinction between short- and long-term repeats, and thus the definition of a measurement session, may not always be obvious or unique for a given trait. 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 reasonable measurement session could then be a single day, and within-day repeats can thus be used to estimate $\sigma^2_e$. If however any fluctuations in body mass are of interest, irrespective of how persistent they are, much shorter measurement sessions, such as seconds or minutes, would be appropriate to ensure that only the purely mechanistic measurement error variance is represented by $\sigma^2_e$.
REPEATED MEASUREMENTS IN THE ANIMAL MODEL

In the following, we show how measurement error can be incorporated in the key tool of quantitative genetics, the animal model, a special type of (generalized) linear mixed model, which is commonly used to decompose the phenotypic variance of a trait into genetic and nongenetic components (Henderson 1976; Lynch and Walsh 1998; Kruuk 2004).

Let us assume that phenotypic measurements of a trait are blurred by measurement error following model (5), and that measurements have been taken both across and within multiple measurement sessions, as indicated in Fig. 1A. Denoting by \( z_{ijk} \) the \( k \)th measurement of individual \( i \) in session \( j \), it is possible to fit a model that decomposes the trait value as

\[
    z_{ijk}^* = \mu + x_{ijk}^T \beta + a_i + id_i + R_{ij} + e_{ijk},
\]

where \( \mu \) is the population intercept, \( \beta \) is a vector of fixed effects, and \( x_{ijk} \) is the vector of covariates for measurement \( k \) in session \( j \) of animal \( i \). The remaining components are the random effects, namely the breeding value \( a_i \) with dependency structure \((a_1, \ldots, a_n)^T \sim N(0, \sigma^2_A \mathbf{A})\), an independent, animal-specific permanent environmental effect \( id_i \sim N(0, \sigma^2_{PE})\), an independent Gaussian residual term \( R_{ij} \sim N(0, \sigma^2_R)\), and an independent error term \( e_{ijk} \sim N(0, \sigma^2_e)\) that absorbs any transient effects captured by the within-session repeats. The dependency structure of the breeding values \( a_i \) is encoded by the additive genetic relatedness matrix \( \mathbf{A} \) (Lynch and Walsh 1998), which is traditionally derived from a pedigree, but can alternatively be calculated from genomic data (Meuwissen et al. 2001; Hill 2014). The model can be further expanded to include more fixed or random effects, such as maternal, nest, or time effects, but we omit such terms here without loss of generality. Importantly, model (7) does not require that all individuals have repeated measurements in each session to obtain an unbiased estimate of the variance components in the presence of measurement error. In fact, even if there are, on average, fewer than two repeated measurements per individual within sessions, it may be possible to separate the error variance from the residual variance, as long as the total number of within-session repeats over all individuals is reasonably large. We will in the following refer to model (7) as the “error-aware” model.

If, however, a trait has not been measured across different time scales (i.e., either only within or only across measurement sessions), not all variance components are estimable. In the first case, when repeats are only taken within a single measurement session for each individual, as depicted in Fig. 1B, an error term can be included in the model, but a permanent environmental effect cannot. The model must then be reduced to

\[
    z_{ik}^* = \mu + x_{ik}^T \beta + a_i + R_i + e_{ik},
\]

thus it is possible to estimate the error variance \( \sigma^2_e \) and to obtain unbiased estimates of \( \sigma^2_A \) and \( h^2 \), whereas the residual variance \( \sigma^2_R \) then also contains the permanent environmental variance. In the second case, when repeated measurements are only available from across different measurement sessions, as illustrated in Fig. 1C, the error variance cannot be estimated. Instead, an animal-specific permanent environmental effect can be added to the model, which is then given as

\[
    z_{ij}^* = \mu + x_{ij}^T \beta + a_i + id_i + R_{ij}
\]

for the measurement in session \( j \) for individual \( i \). Interestingly, this last model mirrors the types of repeats that motivated
quantitative geneticists to isolate $\hat{\sigma}^2_{z^*}$, which may otherwise be confounded not only with $\sigma^2_{\epsilon}$, but also with $\sigma^2_{\lambda}$. This occurs because the repeated measurements across sessions induce an increased within-animal correlation (i.e., a similarity) that may be absorbed by $\sigma^2_{\lambda}$ if not modeled appropriately (Kruuk and Hadfield 2007; Wilson et al. 2010).

**MEASUREMENT ERROR AND 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 $\beta_z$ as the slope of the regression of relative fitness $w$ on the phenotypic trait $z$:

$$w = \alpha + \beta_z \cdot z + \epsilon,$$

with intercept $\alpha$ and residual error vector $\epsilon$. This model can be further extended to account for covariates, such as sex or age. If the phenotype $z$ is measured with error (which may again encompass any irrelevant fluctuations), such that the observed value is $z^* = z + \epsilon$ with error variance $\sigma^2_{\epsilon}$, as in (5), the regression of $w$ against $z^*$ leads to an attenuated version of $\beta_z$ (Fuller 1987; Mitchell-Olds and Shaw 1987; Carroll et al. 2006). Using that $\hat{\beta}_z = \frac{\beta(z, w)}{\beta(z^*, w)} = \frac{\sigma_p(z, w)}{\sigma_p(z^*, w)} + \sigma_{\epsilon_m}^2$, and the assumption that the error in $z^*$ is independent of $w$, simple calculations show that the error-prone estimate of selection is

$$\hat{\beta}_z^* = \frac{\sigma_p(z^*, w)}{\sigma_p(z^*)} = \frac{\sigma_p(z, w)}{\sigma_p(z) + \sigma_{\epsilon_m}^2} \leq \hat{\beta}_z.$$

Hence, the quantity that is estimated is $\beta_z^* = \lambda \beta_z$ with $\lambda = \sigma^2_{\epsilon}(z)/\sigma^2_{\epsilon}(z) + \sigma_{\epsilon_m}^2$, thus $\hat{\beta}_z$ suffers from exactly the same bias as the estimate of heritability (see again Table 1). To obtain an unbiased estimate of selection it may thus often be necessary to account for the error by a suitable error model. Such error-aware model must rely on the same type of short-term repeated measurements as those used in (7) or (8), but with the additional complication that $z$ is now a covariate in a regression model, and no longer the response. To estimate an unbiased version of $\beta_z$, we therefore rely on the interpretation as an error in variables problem for classical measurement error (Fuller 1987; Carroll et al. 2006). To this end, we propose to formulate a Bayesian hierarchical model, because 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 obtain an error-aware model that accounts for error in selection gradients, we need a three-level hierarchical model: the first level is the regression model for selection, and the second level is given by the model of the observed covariate $z^*$ given its true value $z$. Third, a so-called exposure model for the unobserved (true) trait value is required to inform the model about the distribution of $z$, and it seems natural to employ the animal model (9) for this purpose. Again using the notation 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

$$w_{ij} = \alpha + \beta_z z_{ijk} + x_{ij}' \beta + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2_{\epsilon})$$

Selection model (11a)

$$z_{ijk} = z_{ij} + \epsilon_{ijk}, \epsilon_{ijk} \sim N(0, \sigma^2_{\epsilon})$$

Error model (11b)

$$z_{ij} = \mu + x_{ij}' \gamma + a_i + d_i + R_{ij}, R_{ij} \sim N(0, \sigma^2_{\epsilon})$$

Exposure model (11c)

where $w_{ij}$ is the measurement of relative fitness for individual $i$, usually taken only once per individual and having the same value for all measurement sessions $j$. $\beta$ is a vector of fixed effects, $x_{ij}$ is the vector of covariates for animal $i$ in measurement session $j$. $\beta_z$ is the selection gradient, and $\alpha$ and $\epsilon_{ij}$ are, respectively, the intercept and the independent residual term from the linear regression model. The classical independent measurement error term is given by $\epsilon_{ijk}$, and the vector of fixed effects in the exposure model is now denoted as $\gamma$ to discriminate it from $\beta$ in the selection model. This formulation as a hierarchical model gives an unbiased estimate of the selection gradient $\beta_z$, because the lower levels of the model properly account for the error in $z$ by explicitly modeling it. It might be helpful to see that the second and third levels are just a hierarchical representation of model (7). Model (11a-c) can be fitted in a Bayesian setup, see for instance Muff et al. (2015) for a description of the implementation using integrated nested Laplace approximations (INLA) (Rue et al. 2009) via its R interface R-INLA.

Note that the selection model (11a) is formulated here for directional selection. Although the explicit discussion of alternative selection mechanisms, such as stabilizing or disruptive selection, is beyond the scope of the present article, we note that error modeling for these cases is straightforward: The only change is that the linear selection model is replaced by the appropriate alternative, for example by including quadratic or any other kind of nonlinear terms (e.g. Fisher 1930; Lande and Arnold 1983). Moreover, (11a) can be replaced by any other regression model, for example by one that accounts for nonnormality of fitness (see e.g., Morrissey and Sakrejda 2013; Morrissey and Goudie 2016). Similarly, it is conceptually straightforward to replace the Gaussian error and
Table 1. Overview of the effects of measurement error and transient fluctuations (abbreviated as ME) on important quantitative genetic parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Notation</th>
<th>Effect of ME</th>
<th>Biased parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive genetic variance</td>
<td>$\sigma_a^2$</td>
<td>Unbiased</td>
<td>–</td>
</tr>
<tr>
<td>Permanent environmental variance</td>
<td>$\sigma^2_{PE}$</td>
<td>Unbiased</td>
<td>–</td>
</tr>
<tr>
<td>Residual variance</td>
<td>$\sigma^2_r$</td>
<td>Biased</td>
<td>$\sigma^2_r + \sigma^2_e$</td>
</tr>
<tr>
<td>Heritability</td>
<td>$h^2$</td>
<td>Biased</td>
<td>$\lambda h^2$</td>
</tr>
<tr>
<td>Selection gradient</td>
<td>$\beta_z$</td>
<td>Biased</td>
<td>$\lambda \beta_z$</td>
</tr>
<tr>
<td>Phenotypic covariance</td>
<td>$\sigma_P(z, w) = S$</td>
<td>Unbiased</td>
<td>–</td>
</tr>
<tr>
<td>Response to selection (STS)</td>
<td>$\sigma_P(z, w) = R_{STS}$</td>
<td>Unbiased</td>
<td>–</td>
</tr>
<tr>
<td>Response to selection (BE)</td>
<td>$R_{BE}$</td>
<td>Biased</td>
<td>$\lambda R_{BE}$</td>
</tr>
<tr>
<td>Evolvability</td>
<td>$I$</td>
<td>Unbiased</td>
<td>–</td>
</tr>
</tbody>
</table>

The table indicates for each parameter whether it is biased or unbiased. For biased parameters the quantities that are estimated when ignoring transient effects in the quantitative genetic models are given. $\lambda$ is the reliability ratio, defined as $\lambda = \frac{\sigma^2_P}{\sigma^2_P + \sigma^2_{\text{err}}}$. For notation see the main text.

exposure models, if there is reason to believe that the normal assumptions for the error term $e_{ijk}$ or the residual term $R_{ij}$ are unrealistic, for example if $z$ is a count or a binary variable. In fact, equation (10) to estimate selection does not actually assume a specific distribution for $z$, however, the interpretation of $\beta_z$ as a directional selection gradient to predict evolutionary change may be lost for non-Gaussian traits (Lande and Arnold 1983). Finally and importantly, although multivariate selection is not covered in the present article, it is possible to extend the hierarchical model (11a–c) to the multivariate case.

To understand how transient effects affect the estimate of $R_{BE} = h^2 : S$, we must understand how the components $h^2$ and $S$ are affected. We have seen that $h^2 = \lambda h^2$. On the other hand, the selection differential $S = \sigma_P(z^*, w)$ is an unbiased estimate of $\sigma_P(z, w)$, because under the assumption of independence of the error vector $e$ and fitness $w$,

$$\sigma_P(z^*, w) = \sigma_P(z + e, w) = \sigma_P(z, w) + \sigma_P(e, w) = \sigma_P(z, w).$$  (12)

Consequently, the bias in $h^2$ directly propagates to the estimated response to selection, that is, $R_{BE}^* = \lambda R_{BE}$ (Table 1).

The Robertson–Price identity
Response to selection can also be predicted using the secondary theorem of selection. Specifically, the additive genetic covariance of the relative fitness $w$ and the phenotypic trait $z$, $\sigma_P(w, z)$, can be estimated from a bivariate animal model. If interest centers around the evolutionary response of a single trait, the model for the response vector including the (error-prone) trait values $z^*$ and relative fitness values $w$ is bivariate with

$$\begin{bmatrix} z^* \\ w \end{bmatrix} = \mu + X\beta + Da + Zr,$$  (13)

where $\mu$ is the intercept vector, $\beta$ the vector of fixed effects, $X$ the corresponding design matrix, $D$ is the design matrix for the breeding values $a$, and $Z$ is a design matrix for additional random terms $r$. 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

$$
\mathbf{a} = \begin{bmatrix} a(z^*) \\ a(w) \end{bmatrix} \sim \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} \sigma^2_a(z^*) \mathbf{A} & \sigma_a(w, z^*) \mathbf{A} \\ \sigma_a(w, z^*) \mathbf{A} & \sigma^2_a(w) \mathbf{A} \end{bmatrix} \right),
$$

(14)

where \(a(z^*)\) and \(a(w)\) 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 \(\sigma_a(w, z^*)\) is extracted from this covariance matrix. An interesting feature of the additive genetic covariance, and consequently estimates of the response to selection using the STS, is that it is unbiased by independent error in the phenotype. This can be seen by reiterating the exact same argument as in equation (12), but replacing the phenotypic with the genetic covariance.

We confirmed all these theoretical expectations with a simulation study, where we analyzed the effects of measurement error on the estimates of interest by adding error terms with different variances to the phenotypic traits. Details and results of the simulations are given in Appendix S2, whereas the code for their implementation is reported in Appendix S3.

**Empirical Example**

**DATA AND MODELS FOR 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 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 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 (in grams). The dataset contained 3266 mass observations from 917 different voles across nine 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 contain irrelevant within-day fluctuations (eating, defecating, digestive processes), but also unknown pregnancy conditions in females, which cannot reliably be determined in the field. Repeated measurements were available, both recorded within and across different seasons. In each season two to five “trapping sessions” were conducted, which each lasted four consecutive nights. Although this definition of measurement session was based purely on operational aspects driven by the data collection process, we used this time interval to estimate \(\sigma^2_e\). It is arguably possible that four days might be undesirably long, and that variability in such an interval includes more than purely transient effects, but the data did not allow for a finer time resolution. However, to illustrate the importance of the measurement session length, we also repeated all analyses with measurement sessions defined as a calendar month, which is expected to identify a larger (and probably too high) proportion of variance as \(\sigma^2_e\). The number of four-day measurement sessions per individual was on average 3.02 (min = 1, max = 24) with 1.15 (min = 1, max = 3) number of short-term repeats on average, whereas there were 2.37 (min = 1, max = 13) one-month measurement sessions on average, with 1.41 (min = 1, max = 6) short-term repeats per measurement session.

**Heritability**

Bonnet et al. (2017) estimated heritability using an animal model with sex, age, Julian date (JD), squared Julian date, and the two- 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 variances (de Boer and Hoeschele 1993). The breeding value \((a_i)\), the maternal identity \((m_i)\), and the permanent environmental effect explained by the individual identity \((id_i)\) were included as individual-specific random effects.

If no distinction is made between short-term (within measurement session) and long-term (across measurement sessions) repeated measurements, the model that we denote as the naive model is given as

$$
\mathbf{z}_{ijk}^* = \mathbf{\mu} + \mathbf{x}_{ijk}^\top \mathbf{\beta} + a_i + m_i + id_i + R_{jk},
$$

(15)

where \(\mathbf{z}_{ijk}^*\) is the mass of animal \(i\) in measurement session \(j\) for repeat \(k\). This model is prone to underestimate heritability, because it does not separate the variance \(\sigma^2_e\) from the residual variability, and \(\sigma^2_e\) is thus treated as part of the total phenotypic trait variability. To isolate the measurement error variance, the model expansion

$$
\mathbf{z}_{ijk}^* = \mathbf{\mu} + \mathbf{x}_{ijk}^\top \mathbf{\beta} + a_i + m_i + id_i + R_{ij} + e_{ijk},
$$

with \(R_{ij} \sim \mathcal{N}(0, \sigma^2_{Rij})\) and \(e_{ijk} \sim \mathcal{N}(0, \sigma^2_{eijk})\) leads to what we denote here as the error-aware model. Under the assumption that the length of a measurement session was defined in an appropriate way, and that the error obeys model (5), this model yields an unbiased estimate of \(h^2\), calculated as \(\frac{\sigma^2_a}{\sigma^2_a + \sigma^2_e + \sigma^2_{Rij} + \sigma^2_{eijk}}\) (in agreement with Bonnet et al. 2017), where \(\sigma^2_{eijk}\) is explicitly estimated and thus not included in the denominator. Both models were implemented in \texttt{MCMCglmm} (Hadfield 2010) and are reported in Appendix S4. Inverse gamma priors IG(0.01, 0.01), parameterized with shape and rate parameters, were used for all variances in all models,
Table 2. Estimates of quantitative genetic parameters of body mass in snow voles using naive and error-aware models.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \hat{h}^2 )</th>
<th>( \hat{\sigma}^2_A )</th>
<th>( \hat{\sigma}^2_{PE} )</th>
<th>( \hat{\sigma}^2_M )</th>
<th>( \hat{\sigma}^2_R )</th>
<th>( \hat{\sigma}^2_{\text{res}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.14</td>
<td>3.40</td>
<td>6.09</td>
<td>1.16</td>
<td>12.40</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>[0.07, 0.25]</td>
<td>[1.41, 6.15]</td>
<td>[4.33, 8.51]</td>
<td>[0.56, 2.84]</td>
<td>[11.78, 13.21]</td>
<td>–</td>
</tr>
<tr>
<td>Error-aware (four-day measurement session)</td>
<td>0.23</td>
<td>3.97</td>
<td>5.62</td>
<td>1.48</td>
<td>6.58</td>
<td>6.07</td>
</tr>
<tr>
<td></td>
<td>[0.09, 0.33]</td>
<td>[1.46, 6.06]</td>
<td>[3.68, 7.68]</td>
<td>[0.57, 2.73]</td>
<td>[5.76, 7.82]</td>
<td>[5.54, 7.05]</td>
</tr>
<tr>
<td>Error-aware (one-month measurement session)</td>
<td>0.24</td>
<td>3.82</td>
<td>4.78</td>
<td>1.58</td>
<td>5.77</td>
<td>7.91</td>
</tr>
<tr>
<td></td>
<td>[0.10, 0.37]</td>
<td>[1.17, 5.84]</td>
<td>[3.16, 7.21]</td>
<td>[0.61, 2.86]</td>
<td>[4.78, 6.71]</td>
<td>[7.15, 8.38]</td>
</tr>
</tbody>
</table>

The posterior modes of variance components and heritability are given, together with their 95% credible intervals (in brackets).

whereas \( N(0,10^{12}) \) (i.e., default MCMCglmm) priors were given to the fixed effect parameters. Analyses were repeated with varying priors on \( \sigma^2_{\text{res}} \) for a sensitivity check, but results were very robust (results not shown).

Response to selection
Response to selection on body mass was estimated with rLRS using the Breeder’s equation (1) and the secondary theorem of selection (3), both for the naive and the error-aware versions of the model. The naive and error-aware versions of \( R_{\text{BE}} \) were estimated by substituting either the naive \( h^2 \) or the error-aware estimates of \( h^2 \) into the Breeder’s equation, where the selection differential was calculated as the phenotypic covariance between mass and rLRS. On the other hand, \( R_{\text{STS}} \) was estimated from the bivariate animal model, implemented in MCMCglmm using the same fixed and random effects as those in equation (15). Again IG(0.01, 0.01) priors were used for the variance components. No residual component was included for the fitness trait, as suggested by Morrissey et al. (2012), and its error variance was fixed at 0, because no error modeling is required. Appendix S6 contains the respective R code.

SNOW VOLES RESULTS
Heritability
As expected from theory (Table 1), transient effects in the measurements of body mass biased some, but not all, quantitative genetic estimates in our snow vole example (Table 2). The estimates and confidence intervals of the additive genetic variance \( \hat{\sigma}^2_A \), as well as the permanent environmental variance \( \hat{\sigma}^2_{PE} \) and the maternal variance (denoted as \( \hat{\sigma}^2_M \)) were only slightly corrected in the error-aware models. Residual variances, however, were much lower when measurement error was accounted for in the models. The error-aware model separated residual and transient (error) variance so that \( \hat{\sigma}^2_R + \hat{\sigma}^2_{\text{res}} \) corresponded approximately to \( \hat{\sigma}^2_R \) from the naive model. The overestimation of the residual variance resulted in estimates of heritability that were underestimated by nearly 40% when measurement error was ignored (\( \hat{h}^2 = 0.14 \) in the naive model and \( \hat{h}^2 = 0.23 \) in the error-aware model).

As expected, the estimated measurement error variance was larger when a measurement session is defined as a full month (\( \hat{\sigma}^2_{\text{res}} = 7.91 \)) than as a 4-day interval (\( \hat{\sigma}^2_{\text{res}} = 6.07 \); Table 2), whereas \( N(0, 10^{12}) \) priors were given to the fixed effect parameters. Analyses were repeated with varying priors on \( \sigma^2_{\text{res}} \) for a sensitivity check, but results were very robust (results not shown).
because the trait then has more time and opportunity to change. As a consequence, heritability is even slightly higher (\(\hat{h}^2 = 0.24\)) when the longer measurement 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, and not biologically relevant variability of the phenotypic trait. In the case of the mass of a snow vole, most biologists would probably agree that changes in body mass over a one-month measurement session may well be biologically meaningful and relevant (i.e., body fat accumulation, pregnancy in females, etc.), while it is less clear how much of the fluctuations within a 4-day measurement 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, because within-day variance is likely mostly transient, but because the data were not collected with the intention to quantify such effects, within-day repeats were not available in sufficient numbers in our example dataset.

**Selection**

As expected, estimates of selection gradients (\(\hat{\beta}_c\)) obtained with the naive models provided nearly 40% lower estimates of selection than the error-aware model (Table 3). The two measurement session lengths yielded similar results. With and without measurement error modeling, the \(P\) values of the zero-inflated Poisson models confirmed the presence of selection on body mass in snow voles (\(P < 0.001\) in all models).

<table>
<thead>
<tr>
<th>Model</th>
<th>(\hat{\beta}_c)</th>
<th>(P)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.065</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Error-aware (four-day measurement session)</td>
<td>0.104</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Error-aware (one-month measurement session)</td>
<td>0.104</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

For all models, Bayesian \(P\) values were derived from zero-inflated Poisson regressions.

**Response to selection**

In line with theory, estimates of the response to selection using the Breeder’s equation were nearly 40% lower when transient effects were not incorporated in the quantitative genetic models using four-day measurement sessions (\(\hat{R}_{BE} = 0.10\) in the naive model and \(\hat{R}_{BE} = 0.16\) in the error-aware model; Table 4). As in the case of heritability, the one-month measurement session definition resulted in even slightly higher estimates of the response to selection (\(\hat{R}_{BE} = 0.17\)). In contrast, response to selection measured by the secondary theorem of selection \(\hat{R}_{STS}\) did not show evidence of bias, and the error-aware model with a four-day measurement session definition estimated the same value (\(\hat{R}_{STS} = -0.17\)) as the naive model (Table 4). With a one-month measurement session, we obtained a slightly attenuated value (\(\hat{R}_{STS} = -0.14\)), although the difference was small in comparison to the credible intervals (Table 4).

This example illustrates that the Breeder’s equation is generally prone to underestimation of the selection response in real study systems when measurement error in the phenotype is present (Table 1). The results also confirm that estimates for response to selection may differ dramatically between the Breeder’s equation and the secondary theorem of selection. As already noticed by Bonnet et al. (2017), the predicted evolutionary response derived from the Breeder’s equation points in the opposite direction in the snow vole data than the estimate derived from the secondary theorem of selection (e.g., naive estimates \(\hat{R}_{BE} = 0.10\) vs. \(\hat{R}_{STS} = -0.17\), with nonoverlapping credible intervals; Table 4).

**Discussion**

This study addressed the problem of measurement error and transient fluctuations in continuous phenotypic traits in quantitative genetic analyses. We have shown that measurement error and transient fluctuations can lead to substantial bias in estimates of several important quantitative genetic parameters, including heritability, selection gradients, and the response to selection (Table 1). We introduced modeling strategies to obtain unbiased estimates in these parameters in the presence of measurement error and transient fluctuations. These strategies rely on the distinction between variability from stable effects that are part of the biologically relevant phenotypic variability, and transient effects, which are the sum of mechanistic measurement error and biological fluctuations that are considered irrelevant for the selection process. We argue that ignoring the distinction between stable and transient effects may not only lead to an underestimation of the heritability due to inflated estimates of the residual variance, \(\sigma^2_k\), but also to bias in the estimates of selection gradients and the response to selection. Measurements of the same individual repeated at appropriate time scales allow the variance from such transient effects to be partitioned, and thus prevent such bias.

How can repeated measurements be used to prevent an underestimation of heritability, selection, and response to selection, while permanent environment effects are required in quantitative genetic models of repeated measures to avoid an upward bias of \(\sigma^2_k\) and, hence, an overestimation of \(h^2\) (Wilson et al. 2010)? The fact that repeated measurements are used to prevent opposite
biases in heritability estimates makes it apparent that the information content in what is termed “repeated measurements” in both cases is very different. The crucial aspect 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. Repeats taken on the same individual at different life stages (“long-term” repeats, e.g., across what we call measurement sessions here) can be used to separate the animal-specific permanent environmental effect from both genetic and residual variances. On the other hand, repeats taken in temporal vicinity (“short-term” repeats, e.g., within a measurement session) help disentangle any transient from the residual effects. Only by modeling both types of repeats, that is, across different relevant time scales, is it practically feasible to separate all variance components. To do so, the quantitative genetic model for the trait value, typically the animal model, needs extension to three levels of measurement hierarchy (eq. 7): the individual (i), the measurement session (j within i), and the repeat (k within j within i). As highlighted with the snow vole example, it may not always be trivial to determine, in a particular system, an appropriate distinction between short- and long-term repeats, and consequently how to define a measurement session. This decision must be driven by the definition of short-term variation as a variation that is not “seen” by the selection process (see, e.g., Price and Boag 1987, p. 279 for a similar analogy) in contrast to persistent effects that are potentially under selection. This distinction ultimately depends on the trait, on the system under study and on the research question that is asked, because some traits may fluctuate on extremely short time scales (minutes or days), whereas others remain constant across an entire adult’s life.

The application to the snow vole data, where we varied the measurement session length from four days to one month, illustrated that longer measurement sessions automatically capture more variability, that is, the estimated error variance $\hat{\sigma}_e^2$ increased. Consequently, unreasonably long measurement sessions may lead to overcorrected estimates of the parameters of interest. On the other hand, considering measurement sessions that are too short may lead to an insufficient number of within-session repeats, or they may fail to identify transient variability that is biologically irrelevant. This makes clear that a careful definition of measurement session length is important already at the design stage of a study.

If one is uncertain whether repeated measurements capture effects relevant to selection or not, would averaging over repeats result in better estimates of quantitative genetic measures? Averaging methods have been proposed specifically to reduce bias that emerges due to measurement error and transient effects (Carbonaro et al. 2009; Zheng et al. 2016). Although averaging will alleviate bias by reducing the error variance in the mean, it will not eliminate it completely. This can be seen from the fact that averaging over K within-session repeats for all animals and measurement sessions, the variance $\sigma^2_{\text{er}}$ is reduced to $\sigma^2_{\text{er}}/K$, assuming independence of the error term. Unless K is large, $\sigma^2_{\text{er}}$ will not approach zero. Moreover, this practice only works if all animals have the same number of repeats within all measurement sessions, but it will not work in the unbalanced sampling design so common in studies of natural populations.

We approached the problem of measurement error and transient fluctuations by assuming a dichotomous distinction between short- and long-term repeats. This requires determining the temporal scale of measurement sessions, something that is likely highly trait and context specific. An alternative perspective of within-animal repeated measurements could take a continuous view, recalling that repeated measurements are usually correlated, even when taken across long time spans, and that the correlation increases the closer in time the measurements were taken. A more sophisticated model could thus take into account that the residual component in the model changes continuously, and introduce a time-dependent correlation structure instead of simply distinguishing between short- and long-term repeats. Such a model might be beneficial if repeats were not taken in clearly defined measurement sessions, although such a temporal correlation term introduces another level of model complexity, and thus entails other challenges.

It may sometimes not be possible to take multiple measurements on the same individual, or to repeat a measurement within a session. However, it may still be feasible to include an appropriate random effect in the absence of short-term repeats, provided that knowledge about the error variance is available, for example, from previous studies that used the same measurement devices.
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 even fixed) prior on the respective variance component. Such Bayesian models provide error-aware estimates that are equivalent to those illustrated in Table 1, but with the additional advantage that posterior distributions naturally reflect all uncertainty that is present in the parameters, including the uncertainty that is incorporated in the prior distribution of the error variance.

Measurement error and transient fluctuations bias some, but not all quantitative genetic inferences. When \( \sigma^2_{\epsilon_m} > 0 \), the naive estimates of \( h^2 \), \( p_\lambda \), and \( R_{\text{BE}} \) are attenuated by the same factor \( \lambda < 1 \), but other components, such as the selection differential \( S \) or \( R_{\text{STS}} \), are not affected (Table 1). The robustness of the secondary theorem of selection to measurement error can certainly be seen as an advantage over the Breeder’s equation. Nevertheless, the Robertson–Price identity does not model selection explicitly, and thus says little about the selective processes. The Robertson–Price equation can 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). Another quantity that is unaffected by independent transient effects, which we however did not further elaborate on here, is evolvability, defined as the squared coefficient of variation \( I = \sigma^2_A / \bar{z}^2 \), where \( \bar{z} \) denotes the mean phenotypic value (Houle 1992). Evolvability is often used as an alternative to heritability, and is interpreted as the opportunity for selection (Crow 1958). Not only \( \sigma^2_A \), but also \( \bar{z} \) can be consistently estimated using \( z^* \), namely because the expected values \( E[z^*] = E[z] \) due to the independence and zero mean of the error term. For completeness, we added evolvability to Table 1.

A critical assumption of our models was that the error components are independent of the phenotypic trait under study, but also independent of fitness or any covariates in the animal model or the selection model. Although the small changes in \( R_{\text{STS}} \) that we observed in the snow vole application with one-month measurement sessions could be due to pure estimation stochasticity, an alternative interpretation is that the measurement error in the data are not independent of the animal’s fitness. At least two processes could lead to a correlation between the measurement error in mass and fitness in snow voles. First, pregnant females will experience temporally increased body mass, and we expect the positive deviation from the true body mass to be correlated with fitness, because a pregnant animal is likely to have a higher expected number of offspring over its entire lifespan. And second, some of the snow voles were not fully grown when measured, and juveniles are more likely to survive if they keep growing, so that deviations from mean mass over the measurement session period would be nonrandomly associated with life-time fitness.

So far, we have focused on traits that can change relatively quickly throughout the life of an individual, such as body mass, or physiological and behavioral traits. Traits that remain constant after a certain age facilitate the isolation of measurement error, because the residual variance term is then indistinguishable from the error term, given that a permanent environmental (i.e., individual-specific) effect is included in the model. In such a situation it is sufficient to estimate \( \sigma^2_{R} \), which then automatically corresponds to the measurement error variance, whereas \( \sigma^2_{PE} \) captures all the environmental variability. However, not many traits will fit that description. The majority of traits, even seemingly stable traits such as skeletal traits, are in fact variable over time (Price and Grant 1984; Smith et al. 1986).

We have shown that dealing appropriately with measurement error and transient fluctuations of phenotypic traits in quantitative genetic analyses requires the inclusion of additional variance components. Quantitative genetic analyses often differ in the variance components that are included to account for important dependencies in the data (Meffert et al. 2002; Kruuk and Hadfield 2007; Palucci et al. 2007; Hadfield et al. 2013). Besides the importance of separating the right variance components, it has been widely discussed which of the components are to be included in the denominator of heritability estimates, although the focus has been mainly on the proper handling of variances that are captured by the fixed effects (Wilson 2008; de Villemereuil et al. 2018). We hope that our treatment of measurement error in quantitative genetic analyses sparks new discussions of what should be included in the denominator when heritability is calculated.

The methods presented in this article have been developed and implemented for continuous phenotypic traits. Binary, categorical or count traits may also 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 (e.g., Muff et al. 2018). 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). In these cases, it will often not be possible to obtain unbiased estimates of quantitative genetic parameters by adding an error term to the linear predictor as we have done here for continuous traits. Obtaining unbiased estimates of quantitative genetic parameters in the presence of misclassification and miscounting error will require extended modeling strategies, such as hierarchical models with an explicit level for the error process.

We hope that the concepts and methods provided here serve as a useful starting point when estimating quantitative genetics parameters in the presence of measurement error or transient, irrelevant fluctuations in phenotypic traits. The proposed approaches are relatively straightforward to implement, but further generalizations are possible and will hopefully follow in the future.
**AUTHOR CONTRIBUTIONS**

SM and LFK conceived the research idea. EP designed and conducted the simulations and analyses. TB collected and compiled the vole data. EP and SM wrote the manuscript, with inputs from LFK and TB. All authors gave final approval.

**ACKNOWLEDGMENTS**

This work would not have been possible without generous funding from the Faculty of Science and the Forschungskredit of the University of Zurich, grant no. FK-16-097. The snow vole monitoring was supported by the Swiss National Science Foundation project grants 31003A_141110 and 31003A_159462/1 to E. Postma. We thank E. Postma, G. Camenisch, and P. Wandel for data collection and conceptual input. The thoughtful inputs of two anonymous reviewers and the associate editor, M. Morrissey, greatly improved the manuscript. The authors declare they have no competing interests.

**DATA ARCHIVING**

Data has been fully archived in https://doi.org/10.1371/journal.pbio.1002592

**LITERATURE CITED**


Associate Editor: M. Morrissey

Handling Editor: P. Tiffin

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix 1. Supplementary text and figures (pdf).

Appendix 2. Supplementary text and figures for simulation study (pdf).

Appendix 3. R script for the simulation and analysis of pedigree data.

Appendix 4. R script for heritability in snow voles.

Appendix 5. R script for selection in snow voles.