2-23-2021

Most Published Selection Gradients Are Underestimated: Why This Is and How to Fix It

Niels Jeroen Dingemanse
Ludwig-Maximilians-Universitat Munchen, Germany

Yimen G. Araya-Ajoy
Norwegian University of Science and Technology, Norway

David F. Westneat
University of Kentucky, david.westneat@uky.edu

Most Published Selection Gradients Are Underestimated: Why This Is and How to Fix It

Digital Object Identifier (DOI)
https://doi.org/10.1111/evo.14198

Notes/Citation Information
Published in *Evolution*, v. 75, issue 4.

doi:10.1111/evo.14198

Most published selection gradients are underestimated: Why this is and how to fix it

Niels Jeroen Dingemanse,[1,2] Yimen G. Araya-Ajoy,[3] and David F. Westneat[4]

[1]*Department of Biology, Ludwig-Maximilians-Universitat Munchen Department Biologie II, Planegg-Martinsried, Germany*

[2]*E-mail: n.dingemanse@lmu.de*

[3]*Center for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim 7012, Norway*

[4]*Department of Biology, University of Kentucky, Lexington, Kentucky*

Received July 2, 2020
Accepted February 12, 2021

Ecologists and evolutionary biologists routinely estimate selection gradients. Most researchers seek to quantify selection on individual phenotypes, regardless of whether fixed or repeatedly expressed traits are studied. Selection gradients estimated to address such questions are attenuated unless analyses account for measurement error and biological sources of within-individual variation. Estimates of standardized selection gradients published in *Evolution* between 2010 and 2019 were primarily based on traits measured once (59% of 325 estimates). We show that those are attenuated: bias increases with decreasing repeatability but differently for linear versus nonlinear gradients. Others derived individual-mean trait values prior to analyses (41%), typically using few repeats per individual, which does not remove bias. We evaluated three solutions, all requiring repeated measures: (i) correcting gradients derived from classic models using estimates of trait correlations and repeatabilities, (ii) multivariate mixed-effects models, previously used for estimating linear gradients (seven estimates, 2%), which we expand to nonlinear analyses, and (iii) errors-in-variables models that account for within-individual variance, and are rarely used in selection studies. All approaches produced accurate estimates regardless of repeatability and type of gradient, however, errors-in-variables models produced more precise estimates and may thus be preferable.

KEY WORDS: Bias, measurement error, multivariate mixed-modeling, phenotypic selection, plasticity, repeatability.

Quantifying the strength, direction, and shape of selection is of interest to a variety of biological disciplines. In evolutionary biology, estimates of selection are used to predict evolutionary change (Lande 1979; Lande and Arnold 1983), or to understand the adaptive nature of genetic trait integration (Sinervo and Svensson 2002; Roff and Fairbairn 2012). In evolutionary ecology, variation in selection gradients is used to study the ecology of selection (Siepielski et al. 2009), or to test life history theory (Stearns 1992; Nussey et al. 2007), while behavioral ecologists quantify selection to test predictions of optimality models (Krebs and Davies 1997; Westneat and Fox 2010). With rapid environmental change altering patterns of selection in a myriad of ways (Robertson et al. 2013; Santangelo et al. 2018), accurate estimates of selection are critical (Rivkin et al. 2019).

Regression techniques represent the dominant approach to estimate selection since the seminal paper by Lande and Arnold (1983) published nearly four decades ago. The approach consists, in its simplest form, of regressing relative fitness of an individual as a function of its phenotypic value for a variance-standardized trait to derive standardized selection gradients. Expansion of the regression to include multiple traits, quadratic terms, or interactions between traits enables quantification of many forms of selection, including stabilizing, disruptive, and correlational (Lande and Arnold 1983). The unbiased estimation of selection is key to deriving accurate predictions, and understanding the ecological

drivers, of phenotypic evolution (Kingsolver et al. 2001, 2012; Siepielski et al. 2009). Previous studies have highlighted various sources of bias, including sampling error (Knapczyk and Conner 2007; Kingsolver et al. 2012; Morrissey and Hadfield 2012; Ponzi et al. 2018; Videlier et al. 2020), environmental confounds (Rausher 1992; Stinchcombe et al. 2002; Videlier et al. 2020), and mistakes in the calculation of selection gradients (Stinchcombe et al. 2008).

Most researchers seek to quantify selection on individual-specific phenotypes, regardless of whether fixed or repeatedly expressed traits are studied. Here, we focus on the problem that selection gradients estimated to address such questions are attenuated unless analyses (properly) account for measurement error (Ponzi et al. 2018) and—for repeatedly expressed traits—biological sources of within-individual variation. Within-individual variance in fixed traits (e.g., tarsus length in adult birds) results entirely from measurement error; its quantification requires repeated measures. For such traits, the individual's mean trait value across an infinite number of measurements approximates the single (fixed) trait value characterizing the individual (Roff 1997). For repeatedly expressed traits (e.g., behavior), within-individual variation also results from within-individual plasticity. For such traits, individuals are normally assumed to exhibit a norm of reaction, characterized by an average phenotype in the average environmental condition (similar to above) and level of responsiveness (plasticity) to within-individual environmental change (Nussey et al. 2007). Although some explicitly study selection on plasticity (e.g., Nussey et al. 2005; Ramakers et al. 2019), most researchers estimate selection on repeatedly expressed traits to quantify selection on individual-specific (i.e., life-time mean) phenotypes—this is not often stated explicitly. Thus, most studies quantifying selection on individual-mean trait values should aim to account for any form of within-individual variance, whether or not it resulted from a biological (plasticity) or nonbiological (measurement error) process.

Here, we demonstrate that attenuation bias in estimates of standardized selection gradients is inversely related to trait repeatability (see also Ponzi et al. 2018), though differently so for linear versus nonlinear gradients. We detail study designs and statistical approaches enabling unbiased estimation of both types of gradient, the utility of which we verify with simulations. We deem our thesis important because a review of papers estimating standardized selection gradients published in *Evolution* from 2010–2019 inclusive (Supporting Information Text S1 and Table S1) demonstrates that most published studies fail to (properly) control for this form of bias (Table 1). Specifically, most published estimates are based on traits measured once (193 out of 325 estimates; 59%); these are attenuated under realistic residual within-individual error distributions (Ponzi et al. 2018), the extent depending on the type of selection gradient and level of trait repeatability (see section "The Problem"; Table 2). Given that repeatability of most traits generally varies from 0.2 to 0.9 (Bell et al. 2009; Holtmann et al. 2017), bias in estimates and its effect on our ability to interpret patterns of selection is potentially huge. Some have attempted to purge within-individual variance in trait values by deriving individual-mean trait values prior to analyses (132 out of 325 estimates; 41%); most of those calculated individual-means using two to five repeated measures (66 out of these latter 132 estimates; 50%; Table 1). We demonstrate mathematically, and using simulations, that those estimates are also attenuated (see section "The Problem").

Although one solution is to correct estimates of selection gradients based on trait repeatability information (Table 2), two other solutions exist. First, multivariate mixed-effects models provide a general solution for estimating individual-level relationships when predictor and/or response variables are measured with error, or covary differently across hierarchical levels (Browne et al. 2007; Phillimore et al. 2010). Three papers (out of 72; 4%) in our review applied this approach, in all cases to estimate linear selection gradients (Reed et al. 2016; Thomson et al. 2017; Ramakers et al. 2019). Deriving nonlinear selection gradients from multivariate mixed-model approaches requires a simple extension, which we describe below. Second, errors-in-variables (or "measurement error") models have recently been introduced as an alternative solution (Ponzi et al. 2018) and were not employed in any of the studies we reviewed. All three approaches strictly require repeated measures; we use simulations to study bias and precision associated with each approach, for both linear, quadratic, and correlational selection gradient analyses, and for trait repeatability (R) values that are either relatively low (0.3) or high (0.7).

THE PROBLEM

Imagine researchers capturing birds, measuring their tarsus (a component of structural size; a fixed trait), releasing them, and tallying lifetime reproductive success. To estimate the strength of directional selection on tarsus, the researchers can, based on two assumptions, apply two standard transformations to the data (Lande and Arnold 1983). First, they assume lifetime reproductive success (a measure of absolute fitness; W) divided by the population-mean ($\bar{W}$) represents relative fitness (ω). Second, they assume tarsus length (t) divided by the phenotypic standard deviation ($\sqrt{V_{p_t}}$; square-root phenotypic variance (V_p) in t) represents the variance-standardized trait value ($z = t/\sqrt{V_{p_t}}$). Researchers then fit a linear regression, assuming the following true relationships:

$$\omega = \alpha + \beta_1^* z + \varepsilon. \quad (1)$$

Here, α represents the intercept, β_1^* is the standardized linear selection gradient, and ε is the residual variance; throughout,

Table 1. Summary statistics associated with a literature review of papers estimating standardized selection gradients published in *Evolution* from 2010 to 2019 inclusive.

Trait Types	Number of studies	Percentage
Morphology	195	63%
Behavior	38	12%
Life history	31	10%
Physiology	26	8%
Performance	21	7%
Type of fitness metric[1] [number of repeat measurements per individual]		
Lifetime [1]	101	32%
pisodic [1]	195	63%
Episodic [≥2]	15	5%
Number of repeat measurements (*N*) per trait per individual		
N = 1	191	61%
N = 2–5	91	30%
N > 5	29	9%
Analysis based on mean trait values		
No	212	68%
Yes	88	28%
Other[2]	11	4%
Mentioned repeatability		
No	270	87%
Yes[3]	41	13%
Types of selection gradient measured		
Directional only	133	43%
Quadratic	178	57%
All three[4]	107	34%
Statistical approach used to estimate selection gradients		
Multivariate mixed-effects model	Seven estimates in three papers	2%
Regression/LMM/path analysis	297 estimates from 63 papers	95%
Others (Hurdle, Aster models, splines)	Seven estimates from three papers	2%[5]
Mean trait values transformed to zero prior to analysis		
Mean centered	191	61%
PCA- scores	16	5%
Information not provided	82	26%
Trait correlations provided with estimates of correlational selection		
Yes	61	34%
No	117	66%

[1] Number of trait-fitness estimates; traits are listed once regardless of whether multiple types of selection gradients were estimated, but listed twice if both types of fitness metric (lifetime and episodic) were used (*N* = 311 estimates).
[2] Combined traits or data (PCA or BLUPS).
[3] Twenty-two of which provided estimates, ranges, or solely noted traits were repeatable.
[4] Directional, quadratic, and correlational.
[5] We did not investigate how within-individual variance in traits influences estimates from these techniques, although we suspect there will be similar problems.

parameters ignoring potentially biasing effects of within-individual variance are denoted with a star (*). The problem is that tarsus length is not *fully* repeatable because it is measured with error (e.g., Moiron et al. 2019). The fitness effect (β_1^*) of the variance-standardized trait ($z = t/\sqrt{V_{p_t}}$), therefore, does not reflect the true standardized linear selection gradient (β_1). This is because variance due to measurement error (V_{e_t}) makes V_{p_t} an inflated measure of the among-individual variance (V_{i_t}), and so too the standard deviation. Thus, the definition of standardized trait values assumed above was incorrect: trait values were not

Table 2. Attenuation bias in standardized selection gradients for analyses not (fully) accounting for within-individual trait variance: Correcting for this bias requires dividing the estimated selection gradient by the bias. (A) Bias for analyses based on one trait value per trait (t_1 = trait 1, t_2 = trait 2) per individual. Formulae demonstrate a key role for trait repeatability (R_t) and, for correlational gradients, among-individual ($r_{i_{t_1 t_2}}$) and phenotypic trait correlations ($r_{p_{t_1 t_2}}$). (B) Bias for analyses using individual-mean trait values ($\bar{t}$) instead. Bias then varies with the among-individual variance (V_i), the total phenotypic variance among individual-mean trait values ($V_{p_{\bar{t}}} = V_{i_t} + \frac{V_{e_t}}{n}$, where n is the number of repeated measures per individual used to calculate individual means; eq. S5.1) and, for correlational gradients, $r_{i_{t_1 t_2}}$ and the phenotypic correlation between individual-mean trait values ($r_{p_{\bar{t}_1 \bar{t}_2}}$; defined in footnote 2). These formulae simplify to (A) when $n = 1$. Formulae apply to phenotypic selection analyses that assume no residual covariance between traits and fitness. See Supporting Information Texts S2–S5 for mathematical derivations.

	General formula	(**A**) Formula for one trait value per trait per individual	(**B**) Formula for mean of n trait values per trait per individual
Type of gradient	Attenuation bias	Attenuation bias[1]	Attenuation bias[1,2]
Linear (β_1)	$\sqrt{R_t}$	$\sqrt{R_t}$	$\sqrt{\frac{V_{i_t}}{V_{p_{\bar{t}}}}}$
Quadratic (γ_{11})	$\sqrt{R_{t^2}}$	R_t	$\frac{V_{i_t}}{V_{p_{\bar{t}}}}$
Correlational (γ_{12})	$\sqrt{R_{t_1 t_2}}$	$\sqrt{R_{t_1} R_{t_2}} \sqrt{\frac{r^2_{i_{t_1 t_2}}+1}{r^2_{p_{t_1 t_2}}+1}}$	$\sqrt{\frac{V_{i_{t_1}}}{V_{p_{\bar{t}_1}}} \frac{V_{i_{t_2}}}{V_{p_{\bar{t}_2}}}} \sqrt{\frac{r^2_{i_{t_1 t_2}}+1}{r^2_{p_{\bar{t}_1 \bar{t}_2}}+1}}$

[1] For nonlinear gradients, attenuation biases printed here apply solely to mean-centered traits. Equations (S3.11) and (S4.7) describe biases for, respectively, standardized quadratic and correlational selection gradients when traits were not mean-centered.

[2] The equation $r_{p_{\bar{t}_1 \bar{t}_2}} = (Cov_{i_{t_1 t_2}} + \frac{Cove_{t_1 t_2}}{n}) / \sqrt{(V_{i_{t_1}} + \frac{Ve_{t_1}}{n})(V_{i_{t_2}} + \frac{Ve_{t_2}}{n})}$ (eq. S5.3) shows that the phenotypic correlation between individual-mean trait values ($r_{p_{\bar{t}_1 \bar{t}_2}}$) is affected by the residual within-individual covariance ($Cov_{e_{t_1 t_2}}$) and thus not solely shaped by among-individual covariance ($Cov_{i_{t_1 t_2}}$), although increasingly shaped by the latter with increasing sample size per individual (n).

divided by the true among-individual standard deviation ($\sqrt{V_{i_t}}$) but by an upward-biased proxy ($\sqrt{V_{p_t}} = \sqrt{V_{i_t} + V_{e_t}}$). Furthermore, even if standardization had been applied correctly, β_1^* would still not equal β_1 (see also Ponzi et al. 2018). We demonstrate this by first deriving the true unstandardized linear selection gradient (b_1) from the unstandardized linear selection gradient that ignores within-individual variance (b_1^*), which we then standardize to derive β_1.

Parameter b_1^* represents the slope of the regression of t on W:

$$W = \alpha + b_1^* t + \varepsilon. \tag{2}$$

Here, b_1^* represents the total covariance between the trait and fitness ($C_{p_t,\mathrm{W}}$) divided by the total variance in the trait (V_{p_t}), where $C_{p_t,\mathrm{W}}$ represents the summation of the true among-individual ($C_{i_t,\mathrm{W}}$) and residual covariance ($C_{e_t,\mathrm{W}}$) between W and t, and where $V_{p_t} = V_{i_t} + V_{e_t}$:

$$b_1^* = \frac{C_{p_t,\mathrm{W}}}{V_{p_t}} = \frac{C_{i_t,\mathrm{W}} + C_{e_t,\mathrm{W}}}{V_{i_t} + V_{e_t}} = \frac{C_{i_t,\mathrm{W}}}{V_{i_t}} \frac{V_{i_t}}{V_{i_t} + V_{e_t}} + \frac{C_{e_t,\mathrm{W}}}{V_{e_t}} \frac{V_{e_t}}{V_{i_t} + V_{e_t}} = b_1 R_t + b_{1_e}(1 - R_t). \tag{3}$$

Here, $b_1 = \frac{C_{i_t,\mathrm{W}}}{V_{i_t}}$, $\frac{V_{i_t}}{V_{i_t}+V_{e_t}}$ represents the trait's repeatability (R_t) and $\frac{V_{e_t}}{V_{i_t}+V_{e_t}}$ represents $(1 - R_t)$. Thus, b_1^* varies with b_1 as a function of R_t. Mathematically, b_1^* is also affected by the residual (variance) effect of the trait on fitness ($b_{1_e} = \frac{C_{e_t,\mathrm{W}}}{V_{e_t}}$) but if we assume that measurement error is random with respect to fitness, this term is zero (see Discussion section for consequences of violating this assumption). Equation (3) thus simplifies as follows:

$$b_1^* = b_1 R_t. \tag{4}$$

Equation (4) demonstrates the well-known attenuation effect on (standardized) covariances for predictors measured with error (Fuller 1987; Carroll et al. 2006; Adolph and Hardin 2007). β_1 equals the change in relative fitness per standard deviation unit trait (Lande and Arnold 1983), calculable by dividing b_1 by mean fitness ($\bar{\mathrm{W}}$) and by multiplying this fraction by the square-root of the variance in trait values at the focal level of analysis, thus $\beta_1^* = \frac{\sqrt{V_{i_t}+V_{e_t}}}{\bar{\mathrm{W}}} b_1^*$ and $\beta_1 = \frac{\sqrt{V_{i_t}}}{\bar{\mathrm{W}}} b_1$. The relationship between β_1, β_1^*, and R_t is therefore:

$$\beta_1 = \beta_1^* / \sqrt{R_t}\,. \tag{5}$$

For proof, see Supporting Information Text S2. Notably, Ponzi et al. (2018) derived this bias starting with equations where residual variance (ε) is added to standardized trait values (z); bias then equals R_{z^*} (where $z^* = z + \varepsilon$) rather than $\sqrt{R_t}$. We express bias in units of unstandardized trait values because using repeatability (widely reported in the literature; Bell et al. 2009;

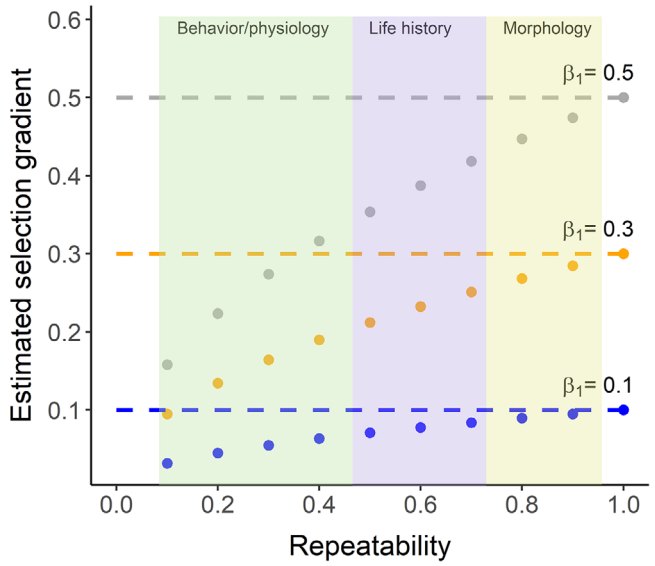

Figure 1. Estimates of standardized linear selection gradients not accounting for within-individual variance (β_1^*) are attenuated by the square-root trait repeatability ($\sqrt{R_t}$). Attenuation varies with trait repeatability (R_t). Dotted lines are true values (β_1). Types of traits differ in typical ranges of *R* (colored zones) and thus in level of bias in estimates of linear selection acting on them.

Holtmann et al. 2017) to correct bias in published estimates of selection gradients is one solution to the problem (see below).

Equation (5) implies all estimates of standardized linear selection gradients are underestimated. Indeed, traits are never *fully* repeatable. In Figure 1, we plot β_1^* as a function of trait repeatability and the true standardized linear selection gradient (β_1) and assigned the typical range of repeatability values (low, moderate, high) to different types of traits to visualize the problem. It shows that, for some types of traits (e.g., behavior, physiology), estimates of directional selection are greatly underestimated when biasing effects of within-individual variance are ignored. Dividing β_1^* by $\sqrt{R_t}$ corrects for this bias, which would obviously require accurate estimates of trait repeatability. One may apply this correction to published estimates. Meta-analytical estimates of trait repeatability are increasingly available in the literature (Holtmann et al. 2017), offering ample opportunities for reanalysis using freely available databases (Kingsolver et al. 2012).

In Supporting Information Texts S3 and S4, we derived attenuation bias for nonlinear standardized selection gradients (γ). Bias in standardized quadratic (stabilizing or disruptive) selection gradients equals the square-root repeatability of squared trait values ($\sqrt{R_{t^2}}$; Supporting Information Text S3); $\sqrt{R_{t^2}}$ varies with trait mean and variance among- and within-individuals (eq. S3.11). Bias in correlational selection gradients equals square-root repeatability of the product of the two focal traits (t_1, t_2) in the analysis ($\sqrt{R_{t_1t_2}}$; Supporting Information Text S4); $\sqrt{R_{t_1t_2}}$ is additionally affected by within- and among-individual trait covariances (eq. S4.7). Attenuation biases in nonlinear gradients thus do not vary solely with trait repeatability. Fortunately, our review implies that most studies mean-center traits prior to selection analysis (Table 1). This removes dependencies on trait means and allows expressing biases in fractions (repeatabilities) and correlations. Attenuation bias ($\sqrt{R_{t^2}}$) in quadratic gradients (γ_{11}) based on analyses of mean-centered traits equals (eq. S3.12):

$$\sqrt{R_{t^2}} = R_t \,. \tag{6}$$

Attenuation bias ($\sqrt{R_{t_1t_2}}$) in correlational gradients (γ_{12}) based on mean-centered traits varies solely with geometric mean trait repeatability ($\sqrt{R_{t_1}R_{t_2}}$), and among-individual ($r_{i_{t_1t_2}}$) and phenotypic ($r_{p_{t_1t_2}}$) trait correlations (eq. S4.9):

$$\sqrt{R_{t_1t_2}} = \sqrt{R_{t_1}R_{t_2}}\sqrt{\frac{r_{i_{t_1t_2}}^2+1}{r_{p_{t_1t_2}}^2+1}}. \tag{7}$$

Here, $r_{p_{t_1t_2}}$ represents the sum of among- ($r_{i_{t_1t_2}}$) and within-individual ($r_{e_{t_1t_2}}$) correlations weighed by $\sqrt{R_{t_1}R_{t_2}}$ (Searle 1961; Dingemanse and Dochtermann 2013):

$$r_{p_{t_1t_2}} = \sqrt{R_{t_1}R_{t_2}}\, r_{i_{t_1t_2}} + \sqrt{\left(1-R_{t_1}\right)\left(1-R_{t_2}\right)} r_{e_{t_1t_2}}. \tag{8}$$

Trait correlations, therefore, do not affect attenuation bias when among- and within-individual correlations are the same, or both zero, because equation (7) then simplifies to $\sqrt{R_{t_1t_2}} = \sqrt{R_{t_1}R_{t_2}}$.

These derivations imply weaker attenuation bias for linear versus nonlinear gradients. For example, when trait correlations do not differ between levels, and for traits mean-centered prior to analyses, attenuation bias for nonlinear gradients equals (geometric mean) trait repeatability (R_t for quadratic; $\sqrt{R_{t_1}R_{t_2}}$ for correlational) rather than its square root ($\sqrt{R_t}$; linear gradients). As above, correcting published estimates of nonlinear gradients is possible but requires estimates of trait means and (co)variances (un-centered traits) or trait repeatabilities and correlations (mean-centered traits), thus information on whether traits were mean-centered. Unfortunately, many studies do not provide all required information (Table 1); correcting published nonlinear estimates may prove challenging. This underlines the need for new studies reporting key descriptive statistics, and above all, applying study designs and statistical approaches avoiding biases altogether.

Few traits have repeatabilities >0.9 and any exhibiting within-individual plasticity often have considerably lower repeatabilities (0.1–0.7; Holtmann et al. 2017). Thus, the attenuation problem is omnipresent and often substantial. Ever since the introduction of the Lande–Arnold approach, researchers have implemented approaches to purge within-individual variation. For example, in his seminal paper on predator-induced correlational selection on stripedness and escape behavior, Brodie (1992)

repeatedly scored each snake's behavior, and used mean values in subsequent selection analyses. This approach is regularly used: 41% of our sample of 325 published values of selection gradients used individual-mean trait values. Taking the mean is, unfortunately, ineffective in purging within-individual error; estimated means only approximate true means when sample sizes approach infinity (Roff 1997).

Specifically, the phenotypic variance among individual-mean trait values ($V_{p_{\bar{i}}}$) approximates the sum of V_{i_t} plus $\frac{V_{e_t}}{n}$ (eq. S5.1), where n represents the number of observations per individual used to calculate individual means (Snijders and Bosker 1999). In Supporting Information Text S5, we make use of this (and other) approximation(s) to derive attenuation bias for selection analyses using individual-mean trait values (Table 2). These derivations imply many repeats are required per individual (n) to fully purge bias caused by effects of within-individual variation, particularly for highly labile traits. For example, when trait repeatability is 0.3, an underestimation of the true standardized linear selection gradient by almost 10% occurs even in the unlikely scenario where individual means were based on 10 replicate measurements per individual. In Table 1, we show that most studies used smaller numbers of measurements per individual (50% of 132 published values used $\leq$5) or failed to provide this information (35 of 132 estimates where $\geq$2 measures were taken; 27%). For 282 of 325 (87%) estimates, trait repeatability was not mentioned, and for only 21 estimates (6%) a repeatability value was given. Correcting published estimates of standardized gradients based on mean trait values will thus also be challenging.

TWO SOLUTIONS

The unbiased estimation of selection gradients requires the partitioning of variances in trait values within versus among individuals, and the estimation of among-individual relationships between traits and fitness. We illustrate the idea by further developing our example of researchers measuring tarsus (t) and lifetime reproductive success (W). Each bird's tarsus was measured repeatedly as part of the study design to estimate—and statistically control for—measurement error. We discuss two types of statistical model proposed to achieve this aim; both strictly require repeated measures data.

Multivariate Mixed-Effects Models

Multivariate mixed-effects models have previously been introduced to estimate linear selection gradients (Morrissey et al. 2010, 2012). The multivariate mixed-effects model offers a two-step solution that starts with estimating among-individual (co)variances between unstandardized trait(s) (t) and absolute fitness (W) from repeated measures data, followed by calculating standardized gradients based on estimated variance components. One can achieve this by fitting the trait (t) and absolute fitness (W) as two responses into a bivariate mixed-effects model with random intercepts for individual identity, resulting in the following phenotypic equation:

$$\begin{bmatrix} t_{hi} \\ W_{hi} \end{bmatrix} = \beta_0 + I_i + e_{hi}. \tag{9}$$

Here, each observation of each response is modeled as a population-mean intercept (β_0) plus the individual's deviation from the population-mean ($+I_i$) plus a residual within-individual error ($+e_{hi}$). Subscripts distinguish between observations (h) and individuals (i). This simplest of bivariate mixed models can be extended by including additional fixed and random effects, ignored here for simplicity. We assume random intercepts (Ω_I) and residuals (Ω_e) follow a multivariate normal distribution (MVN):

$$\begin{bmatrix} I_t \\ I_{\mathrm{W}} \end{bmatrix} \sim \mathrm{MVN}\,(0, \Omega_I) \; : \; \begin{bmatrix} V_{i_t} & C_{i_t,\mathrm{W}} \\ C_{i_t,\mathrm{W}} & V_{i\mathrm{W}} \end{bmatrix}$$
$$\begin{bmatrix} e_t \\ e_{\mathrm{W}} \end{bmatrix} \sim \mathrm{MVN}\,(0, \Omega_e) \; : \; \begin{bmatrix} V_{e_t} & C_{e_t,\mathrm{W}} \\ C_{e_t,\mathrm{W}} & V_{e\mathrm{W}} \end{bmatrix}. \tag{10}$$

Equation (10) may be usefully applied to situations where fitness is episodic and thus repeatedly measured. This bivariate formulation assumes that variances (V) and covariances (C) between the trait and fitness exist both within (e) and among (i) individuals. In the special situation described here (with a single integrative fitness metric per individual), the second term of the equation may be simplified to $[e_t] \sim N(0, \Omega_e) \; : \; [V_{e\mathrm{W}}]$ (Supporting Information Text S8). Other error distributions may be applied to different types of traits. The true standardized linear selection gradient (β_1) detailed in section "The Problem" equals:

$$\beta_1 = \frac{C_{i_t,\mathrm{W}}}{V_{i_t}} \frac{\sqrt{V_{i_t}}}{\beta_{0_{\mathrm{W}}}}. \tag{11}$$

The standardized linear selection gradient (β_1) is thus calculated by multiplying the unstandardized linear gradient ($b_1 = \frac{C_{i_t,\mathrm{W}}}{V_{i_t}}$) by the standard deviation in among-individual trait values ($\sqrt{V_{i_t}}$) and by dividing it by the intercept for fitness ($\beta_{0_{\mathrm{W}}}$); this latter parameter represents mean absolute fitness ($\bar{W}$) in formulations like eq. (9) (where the population-mean intercept is the only fixed effect). Performing these standardizations after rather than before model fitting allows accounting for uncertainty in proxies of means ($\beta_{0_{\mathrm{W}}}$) and variances ($\sqrt{V_{i_t}}$) used to estimate standardized gradients, and thus avoids compounding of estimation error (Hadfield et al. 2010; Houslay and Wilson 2017).

In Supporting Information Text S6, we introduce an extension to estimate nonlinear selection gradients from multivariate mixed-effects models. We propose that unattenuated quadratic selection gradients (γ_{11}) may be acquired by fitting the squared term of the trait (t_{hi}^2) as an additional response (eq. S6.2), or

product of two focal traits (t_1t_2) for correlational selection gradients (γ_{12}). Information embedded in Ω_I is then extracted to calculate unstandardized selection gradients (eq. S6.3); importantly, these calculations differ from the linear scenario because nonlinear selection gradients represent partial regression coefficients (Lande and Arnold 1983), which may be derived by multiplying the inverse matrix of the predictors by the covariance between the traits and fitness (eq. S6.4; for R-code, see Supporting Information Text S8 and future updates on https://github.com/YimenAraya-Ajoy/SelectionBias). Furthermore, formulae for standardizing unstandardized quadratic ($b_{11}\rightarrow\gamma_{11}$; eq. S6.5) and correlational ($b_{12}\rightarrow\gamma_{12}$; eq. S6.7) selection gradients differ from those used to standardize linear gradients ($b_1\rightarrow\beta_1$; eq. 11). Standardized quadratic (γ_{11}; eq. S6.5) versus correlational selection gradients (γ_{12}; eq. S6.7) are, respectively, calculated as (Supporting Information Text S6):

$$\gamma_{11} = 2b_{11}\frac{\sqrt{V_{i_{t^2}}}}{\beta_{0w}}, \tag{12}$$

$$\gamma_{12} = b_{12}\frac{\sqrt{V_{i_{t_1t_2}}}}{\beta_{0w}}. \tag{13}$$

In Supporting Information Texts S3 and S4, we detail how among-individual variances in squared terms ($V_{i_{t^2}}$; eq. S3.10) and products ($V_{i_{t_1t_2}}$; eq. S4.5) can be calculated from trait means and (co)variances.

Errors-in-Variables Models

Errors-in-variables models offer an alternative solution for acquiring unbiased estimates of selection gradients when individual-specific traits values are measured with error. Those models have been called "measurement error" models when introduced to estimate selection gradients (Ponzi et al. 2018); we use the term errors-in-variables models throughout because we apply them here to control for both methodological (measurement error) and biological (phenotypic plasticity) sources of within-individual variance. Compared to multivariate mixed-effects models, errors-in-variables models resolve the problem fundamentally differently: they jointly estimate the expected trait value for each individual as well as its relationship with fitness, which can be described using the following two equations:

$$\begin{aligned} t_{hi} &= \beta_{0_t} + I_{i_t} + e_{hi_t} \\ W_i &= \beta_{0w} + b_1 I_{i_t} + e_{iw}. \end{aligned} \tag{14}$$

Here, the first equation is akin to a univariate mixed-effects model, whereas the second to a linear regression (with unstandardized data) performed at the among-individual level. The standardized gradient (β_1) is then calculated using eq. (11) as above. The example here assumes a simple scenario where within-individual variance results from a single process (e.g., only measurement error); more complex scenarios may also be accommodated (Ponzi et al. 2018). Extending such errors-in-variables models to estimate nonlinear selection requires adding further terms to the second equation (e.g., $I_{i_t}^2$ for modeling quadratic effects). Importantly, β_{0w} (eq. 14) does not represent population-mean fitness ($\bar{W}$) when nonlinear terms are added; this complicates the calculation of standardized gradients; we advise fitting relative (ω_i) instead of absolute (W_i) fitness as a pragmatic solution. See our worked example (Supporting Information Texts S7 and S8) for the subsequent calculation of standardized gradients.

Accuracy and Precision of Each Solution

We used simulations to assess, first, whether classic approaches produced attenuated estimates and second, whether proposed solutions (correcting traditional estimates with measures of repeatability, multivariate mixed-models, and errors-in-variables models) adequately address the problem. We compared systematic error (inaccuracy) and random error (imprecision) across the proposed solutions. We summarize the simulation approach here; we fully describe the approach in Supporting Information Text S7, and provide R-code in Supporting Information Text S8 and on Github (https://github.com/YimenAraya-Ajoy/SelectionBias).

We simulated data assuming a given (linear or nonlinear) relationship between (a) trait(s) and fitness, and then drew three observations per trait per individual; to each observation, we added within-individual variation to generate a target trait repeatability (0.3 or 0.7). Following the generation of the full dataset (n = 800 individuals), we generated two subsets. The first contained one randomly drawn trait value (of the three produced) per individual. The second contained one mean value per individual calculated over all three observations. We then used either classic regression approaches (two subsets), multivariate mixed-effects models (full dataset), or errors-in-variables models (full dataset), to estimate standardized selection gradients. For simplicity, we assumed mean-centered traits, and for correlational selection analyses, zero trait correlations. Under such conditions, all expected attenuation biases were calculable by estimating among- and within-individual trait variances (see Table 2 and Supporting Information Text S8), which we calculated by fitting univariate mixed-effects models (with individual intercepts) to the full dataset. We then calculated "corrected" values by dividing the estimated gradients by the expected attenuation bias. For each (type of) estimate produced, we calculated bias (i.e., inaccuracy or systematic error) as the difference between the estimated (β_1) minus true ($\hat{\beta}_1$) standardized gradient, divided by the true gradient (i.e., $(\beta_1 - \hat{\beta}_1)/\hat{\beta}_1$ for linear gradients). This produced a percentage (upward or downward) bias. For both levels

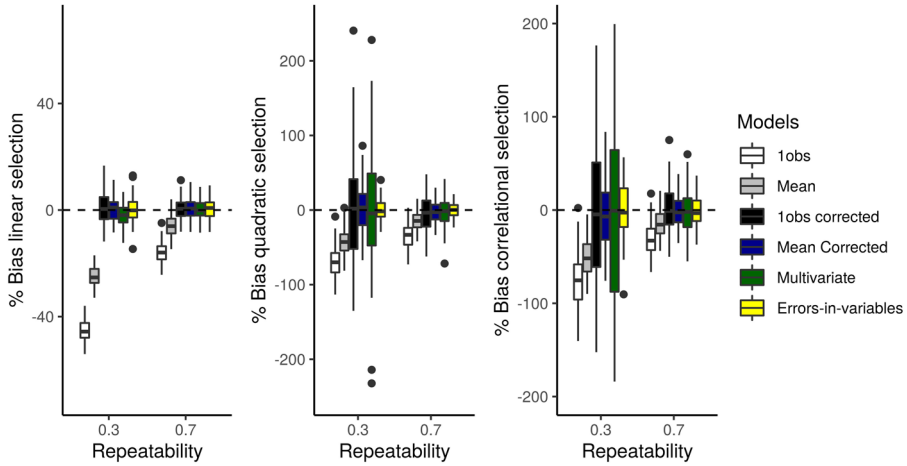

Figure 2. 2Box plots of percent bias in estimates derived from 100 replicate (A) linear, (B) quadratic, or (C) correlational selection analyses. White versus gray boxes indicate results from traditional regression analyses based on a single versus mean (of three) trait value(s) per individual. Black and blue boxes are results from traditional regression analyses based on, respectively, a single versus mean (of three) trait value(s) per individual to which we applied corrections based on predicted bias. Green boxes are results from multivariate mixed-effects models and yellow boxes represent the errors-in-variables models. We estimated bias as the difference between the observed minus simulated standardized selection gradient divided by the simulated standardized selection gradient. Interquartile ranges are inversely related to level of imprecision associated with the given scenario.

of repeatability, we created 100 datasets for linear, quadratic, and correlational selection scenarios. We used the variance among the 100 datasets within a given scenario to assess the expected uncertainty in parameter estimates, in the same way as parametric bootstrapping provides a measure of uncertainty for parameter estimates.

Our simulations showed estimates of linear selection gradients were *always* biased downward when based on traditional approaches, regardless of whether single or individual-means (over three observations) were used. As expected, attenuation bias was less severe for the latter (Fig. 2A), and decreased with increasing trait repeatability. Percentage bias was -45.0% ($R = 0.3$) versus -16.0% ($R = 0.7$) for the single-trait value model, and -24.9% versus -6.0% for the mean-trait value model. These biases disappeared after applying corrections based on predicted bias (single-trait value model: $R = 0.3$: 0.9%, $R = 0.7$: 0.5%); mean-trait value model: $R = 0.3$: 0.5 %, $R = 0.7$: 0.6%). The bivariate mixed-effects model produced accurate estimates (Fig. 2A), both for low ($R = 0.3$; 2.2%) and high ($R = 0.7$; 0.7%) values of repeatability; the same was true for errors-in-variables models (0.3% for $R = 0.3$; 0.4% for $R = 0.7$). Precision was not affected by choice of analytical approach or level of repeatability (Supporting Information Table S7). This is apparent in Figure 2A, where interquartiles (colored ranges) do not vary among box plots.

Simulations applied to quadratic (Fig. 2B) or correlational (Fig. 2C) gradients showed similar patterns of attenuation as above (Fig. 2A), though as expected, nonlinear selection gradients were attenuated more (note differences in *y*-axis scaling across Fig. 2A–C). Corrections were effective in removing bias. Multivariate mixed-effects models produced unbiased estimates as did errors-in-variables models (Supporting Information Table S7; Fig. 2A and B). When repeatability was low, levels of imprecision differed among scenarios (Supporting Information Table S7). Multivariate mixed-effects models then produced relatively imprecise estimates for nonlinear selection gradients. This is evident from comparing interquartile ranges among box plots in Figure 2B and C. Errors-in-variables models were the exception, producing precise estimates regardless of level of repeatability.

GUIDE FOR EMPIRICISTS

How might empiricists go about acquiring unbiased estimates? Our mathematical derivations imply three potential strategies. First, researchers can use a three-step approach by (i) applying classic regression analyses to calculate standardized selection gradients (based on single or individual-mean trait values), (ii) applying mixed-effects models to repeated measures of traits to estimate the (co)variance components required to calculate the expected attenuation bias (using the formulae in Table 2), and (iii) dividing the estimates of standardized selection gradients by their expected attenuation biases. Second, researchers may use multivariate mixed-effects (Morrissey et al. 2010, 2012) or, third, errors-in-variables models (Ponzi et al. 2018). Our simulations imply that multivariate and errors-in-variables models both function appropriately when applied to simple scenarios (linear

selection). Applying classic approaches followed by corrections may, by contrast, often produce estimates dependent on unknown assumptions that may not hold.

An important benefit of errors-in-variables models is that they can produce both accurate and precise estimates under a range of repeatabilities and for all types of selection gradients, although we note that our simulations addressed a limited set of (ideal) conditions. Errors-in-variables models can also usefully consider more complex—yet conceivable—biological scenarios where the residual within-individual variance in one trait is conditional on that of another. For example, measurement error in behavior may differ between large and small animals. This issue is important because it can result in "reverse attenuation" (i.e., overestimation rather than underestimation) in nonlinear selection models (Muff and Keller 2015). Extensions of multivariate mixed-effects models incorporating heterogeneous variance structures (Cleasby et al. 2015) may (partly) achieve the same aim. Importantly, errors-in-variables models allow formulating statistical hypotheses closely matching hypothesized methodological and biological processes. These models are so flexible because observed phenotypes are modeled by formulating distinct equations (e.g., "error" versus "exposure" parts), each with its own distributional assumptions (e.g., binomial versus Gaussian).

Our simulations overall suggest that errors-in-variables models are the preferred approach, certainly when modeling nonlinear selection on traits with low repeatability. A potential concern is that relatively sophisticated programming skills are required and that relatively few software packages are (currently) available for fitting such models. A multivariate mixed-effects modeling approach may therefore also represent a pragmatic solution. Nevertheless, researchers should best invest in learning statistical tools that enable sufficient flexibility to appropriately model the hypothesized data generation processes and thus produce unbiased estimates of selection.

Solutions for propagating uncertainty in estimated parameters are required for all approaches. For multivariate mixed-effects and errors-in-variables models, this may readily be achieved by fitting them in a Bayesian framework (Hadfield 2010; Houslay and Wilson 2017). Bayesian models produce posterior distributions of each parameter, which can be taken to estimate uncertainty associated with derived parameters (e.g., standardized gradients). Similar solutions for the three-step approach associated with classic regression analyses are not obvious.

Discussion

Accurate estimates of selection are crucial for a variety of evolutionary questions. Failure to appreciate that traits are not fully repeatable will result in biased selection gradients (Ponzi et al. 2018). This bias will often come in the form of attenuation (this paper) but "reverse attenuation" (Muff and Keller 2015) may also occur (detailed below). Attenuation bias is arguably problematic regardless of the source of within-individual variation (measurement error or within-individual plasticity). The magnitude of this form of bias differs between types of gradient (Table 2); most papers published in *Evolution* over the past 10 years—presumably a representative sample—fail to appropriately control for it. Our mathematical derivations and literature review imply that many meta-analyses (e.g., Kingsolver et al. 2001, 2012) are based on downward-biased estimates, particularly for nonlinear gradients, and more so for traits with low repeatabilities. This warrants reconsideration of meta-analytical conclusions, by applying corrections to published estimates of selection gradients, re-analyses of repeated measures datasets using more appropriate statistical models, or study-wide adjustments based on meta-analytical estimates of trait repeatability stratified per trait type. We suggest the wide use of the standard approach developed by Lande and Arnold (1983) has resulted in an underestimation of selection, and that new studies should use repeated measures data (regardless of the nature of the trait) and errors-in-variables (or multivariate mixed-models) to acquire unbiased estimates.

Our mathematical derivations suggest relatively straightforward relationships between biased and true selection gradients that vary either with trait repeatability or with its square root (at least for mean-centered data; Table 2). We note that the biases (and how they affect the shape of selection surfaces, discussed below) apply solely when residual covariances (C_e) between traits and fitness are zero. For phenotypic selection analyses based on a single integrative fitness measure per individual (e.g., lifetime reproductive success), this assumption is defendable when one can argue that residual variance in integrative fitness reflects measurement error, and that measurement error should not covary between traits and fitness (because they were determined separately). By contrast, residual covariances are arguably more likely to exist when episodic measures of fitness—like annual reproductive success in species breeding multiple years—are used instead, a common practice in published analyses (Kingsolver et al. 2001, Kingsolver et al. 2012). This is because traits and episodic fitness measures may exhibit within-individual plasticity in response to the same environmental factors. Estimates of selection gradients can then be biased in any direction, depending on whether and how effects of traits on episodic fitness differed within- versus among-individuals. Specifically, there will be no bias caused by failure to account for trait repeatability when among-individual effects of traits on fitness (b_1) do not differ from within-individual effects (b_{1_e}). This can be seen in eq. (3), where b_1^* then equals the true gradient b_1. Indeed, bias occurs only when associations between responses and predictors are underpinned by processes

differing across levels (van Noordwijk and de Jong 1986; Reznick et al. 2000; van de Pol and Wright 2009), such as selection versus measurement error (Dingemanse and Dochtermann 2013), or selection versus plasticity (also called "environmental covariance"; Schlichting 1989; Rausher 1992; Stinchcombe et al. 2002). This is probably the rule than the exception and means that re-analyses using errors-in-variables or multivariate mixed-models may often be required for published estimates of selection inferred from repeated measures of episodic fitness. This makes our call for new studies applying repeated measures designs and statistical approaches avoiding this form of bias even more pressing.

Our analyses imply that bias in selection gradients caused by ignoring within-individual variance may also bias conclusions regarding the ecology of selection. This will particularly be the case when the same ecological factor affects patterns of selection and trait repeatability in concert. Such ecological patterns of covariance between repeatability (or heritability) and selection have been repeatedly demonstrated in wild birds (Brommer et al. 2008; Husby et al. 2011; Nicolaus et al. 2013; Abbey-Lee and Dingemanse 2019). Previous research implied that patterns of selection are stronger in ecological conditions where traits are more repeatable or heritable, potentially speeding up microevolutionary change (Husby et al. 2011; but see Ramakers et al. 2018). Our mathematical derivations demonstrate that such patterns of ecological covariance also result from biases described in this paper (Figure 1; Table 2). Morrissey and Hadfield (2012) implied that an appearance of fluctuating selection where none exists may result from sampling error; our analyses suggest that ecological variation in selection can also be spurious due to unaccounted ecological variation in trait repeatability. Similarly, traits under stronger correlational selection are also more strongly genetically correlated (Roff and Fairbairn 2012), suggesting adaptive evolution of trait correlations. Again, patterns of attenuation bias could also produce such relationships: attenuation bias in correlational selection gradients ($\sqrt{R_{t_1t_2}}$) is inversely related to the strength of the among-individual trait correlation (eq. 7). These examples thereby illustrate a myriad of ways by which unaccounted within-individual variance can result in the appearance of adaptive ecological variation in selection where none exists. Along the same lines, scenarios may also be conceived where ecological variation in selection is masked instead.

Attenuation bias attributable to within-individual variance will also affect conclusions qualitatively. This is because bias differs mathematically between linear ($\sqrt{R_t}$), quadratic ($\sqrt{R_{t^2}}$), and correlational ($\sqrt{R_{t_1t_2}}$) selection gradients (Table 2). We give two examples here. First, attenuation bias can spill over to bias estimates of optimal trait values, and second to bias the shape of correlational selection surfaces. Optimal trait values estimated from quadratic selection analyses are of interest in the context of stabilizing selection. The optimal trait value represents the trait value at the inflection point of the parabola, which equals the linear slope divided by two times the quadratic slope ($\frac{-\beta_1}{2\gamma_{11}}$) (Bronshtein et al. 2015). Because within-individual variability biases linear versus quadratic selection gradients differently (Table 2), estimates of optimal trait values in stabilizing selection scenarios are also affected. For example, for mean-centered traits, the trait value at the parabolic peak is overestimated by a factor $\sqrt{R_t}$ (eq. S3.16). This mismatch is relevant for fields like behavioral ecology that focus on highly labile traits like behavior and routinely ask whether observed trait means match those predicted by optimality theory (Westneat and Fox 2010; Davies et al. 2012). For example, various adaptive explanations have been proposed for why passerines produce smaller clutches than expected according to predictions derived from phenotypic selection analyses (Davies et al. 2012). Because avian clutch size is only moderately repeatable (e.g., Browne et al. 2007), overestimation of the adaptive peak due to failure to account for within-individual variation may thus offer a viable alternative to adaptive explanations.

Attenuation bias can also affect the shape of complex selection surfaces, as commonly derived from correlational selection analyses (Brodie et al. 1995). Surface shape varies with the ratio of the product of the quadratic selection gradients of two focal traits over the square of their correlational selection gradient (i.e., $\gamma_{11}\gamma_{22}/\gamma_{12}^2$), which describes a saddle-shaped fitness surface when below one (assuming γ_{11} and γ_{22} are both negative) but a fitness peak when above one (Phillips and Arnold 1989). In Supporting Information Text S4, we show that the true ratio ($\frac{\gamma_{11}\gamma_{22}}{\gamma_{12}^2}$) equals the ratio derived when one ignores within-individual error ($\frac{\gamma^*_{11}\gamma^*_{22}}{\gamma^*_{12}}$) and that for mean-centered traits (regardless of trait repeatabilities) the fitness surface is unbiased provided trait correlations do not differ between hierarchical levels (e.g., because they are zero). By contrast, the fitness surface is estimated with bias in a conceivable scenario (Dochtermann 2011; Niemelä and Dingemanse 2018; but see Brommer and Class 2017) where among-individual correlations are tighter than overall phenotypic correlations (eq. S4.11), for example, because within-individual variation resulted entirely from measurement error but measurement errors were uncorrelated between traits. Attenuation bias would then make finding saddle-shaped fitness surfaces more likely (Supporting Information Text S4). This problem increases with increasing values of among-individual ($r_{i_{t_1t_2}}$) relative to phenotypic ($r_{p_{t_1t_2}}$) correlations (eq. 7). The occurrence of within-individual variation therefore comes with a large number of (previously unanticipated) consequences with far-reaching consequences.

Applying our proposed approach therefore requires prudent decisions regarding study design and data analyses. First, avoiding repeatability-related biases in selection gradients requires the collection of repeated measures using sampling designs that avoid confounding within- and among-individual associations

(Araya-Ajoy et al. 2015; Ponzi et al. 2018; Mitchell et al. 2019; Westneat et al. 2020). Specifically, inflated estimates of among-individual variance in traits can occur when environmental conditions eliciting reversible plasticity within-individuals are themselves repeatable among individuals (Dingemanse et al. 2010; Westneat et al. 2011, 2020). Ensuring that repeated measures are sufficiently spaced over the lifetime of the individual might help mitigate inflating effects of autocorrelations on trait repeatability (Araya-Ajoy et al. 2015; Allegue et al. 2017; Niemelä and Dingemanse 2017; Ponzi et al. 2018; Mitchell et al. 2019). Second, data analysis strategies should not blindly follow suggestions made here, but be adjusted to specific need. For example, the estimation of age-dependency of selection, or selection on reaction norms, will require modifications of the framework. The meaning of estimated gradients depends also on whether additional fixed or random effects are fitted (e.g., De Lisle and Svensson 2017), an issue also widely applicable to quantitative genetics (Kruuk 2004; Nussey et al. 2007; Hadfield et al. 2010). Continuing the bird example, one might statistically control for size dimorphism by fitting sex as a fixed effect. Assuming lack of sex specificity in Ω_I, application of eq. (9) would then retrieve the standardized linear selection gradient for the reference sex. This is because β_{0_W} reflects mean fitness ($\bar{W}$) of the reference category. The standardized selection gradient for the opposite sex would be calculated by adding the sex-difference in mean fitness (β_{sex_W}) to β_{0_W} in eq. (9). Researchers might also be interested in estimating mean-standardized rather than variance-standardized selection gradients (Houle 1992; Matsumura et al. 2012). If standardizations are applied after rather than before model fitting, this would only require subtle changes in the exact elements used to perform downstream calculations.

Finally, our mathematical derivations (Supporting Information Texts S2–S5) show that the common practice of mean-centering traits prior to analyses, simplifies formulae for attenuation bias in selection analyses not acknowledging within-individual variation. For example, with centered data, the bias in quadratic selection gradients ($\sqrt{R_{t^2}}$) then equals trait repeatability (R_t; eq. 6). Similarly, centering makes bias in correlational selection gradients ($\sqrt{R_{t_1 t_2}}$) vary with geometric mean repeatabilities ($\sqrt{R_{t_1} R_{t_2}}$) and trait correlations (eq. 7). Centering thus makes using classic regression analyses followed by correcting biased estimates easier. The associated overestimation of optimal trait values in the presence of stabilizing selection then also conveniently equals $\sqrt{R_t}$ (eq. S3.16). Mean-centering may also strategically be applied when using multivariate mixed-effects models, because it facilitates model convergence, and biological interpretation of patterns of stabilizing selection (Supporting Information Fig. S3).

In conclusion, our paper highlights the fundamental role of trait repeatability in producing biases in estimates of phenotypic selection gradient analyses. We discuss the possibility of correcting published estimates but highlight that such fixes often may not be applicable, particularly when applied to nonlinear selection gradients. This clarifies that we should do better. We should start using study designs where we always measure traits repeatedly rather than assuming that the level of repeatability does not matter. Only then are we able to apply statistics such as errors-in-variables or multivariate mixed-effects models that enable controlling for biasing within-individual effects. Importantly, our simulations imply that there may not be a single optimal approach; for example, the multivariate mixed-effects models sometimes produce relatively imprecise estimates compared to alternative approaches. Studies using pedigree information and multivariate animal models to estimate genetic gradients, notably, already correct for the highlighted form of bias by partitioning genetic from residual (environmental) variation. Applying any of these two approaches will not only produce better estimates but also more accurate conclusions about the ecology of natural selection.

AUTHOR CONTRIBUTIONS

The idea was conceived by NJD and DFW; NJD developed the math (with major input from YA-A) while birdwatching with DFW; DFW performed the review, and YA-A the simulations. NJD drafted the paper with major input from DFW and YA-A.

ACKNOWLEDGMENTS

We thank Dirk Metzler, Shinichi Nakagawa, Raphael Royauté, Anne Rutten, and Alastair Wilson. NJD was supported by the German Science Foundation (grant no. DI 1694/1-1), YA-A by the Research Council of Norway (Centres of Excellence funding scheme; grant no. 223257), and DFW by the U.S. National Science Foundation (grant no. IOS-1656212) and the University of Kentucky.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA ARCHIVING

All associated data result from data simulations. Statistical scripts are provided as part of the Supporting Information Material, therefore no data are provided.

LITERATURE CITED

Abbey-Lee, R. N., and N. J. Dingemanse. 2019. Adaptive individual variation in phenological responses to perceived predation levels. Nat. Commun. 10:1601.

Adolph, S. C., and J. S. Hardin. 2007. Estimating phenotypic correlations: correcting for bias due to intraindividual variability. Funct. Ecol. 21:178–184.

Allegue, H., Y. G. Araya-Ajoy, N. J. Dingemanse, N. A. Dochtermann, L. Z. Garamszegi, S. Nakagawa, D. Réale, H. Schielzeth, and D. F. Westneat. 2017. Statistical Quantification of Individiual Differences (SQuID): an educational and statistical tool for understanding multilevel phenotypic data in linear mixed models. Methods Ecol. Evol. 8:257–267.

Araya-Ajoy, Y. G., K. J. Mathot, and N. J. Dingemanse. 2015. An approach to estimate short-term, long-term and reaction norm repeatability. Methods Ecol. Evol. 6:1462–1473.

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

Brodie, E. D. 1992. Correlational selection for color pattern and antipredator behavior in the garter snake *Thamnophis ordinoides*. Evolution 46:1284–1298.

Brodie, E. D., A. J. Moore, and F. J. Janzen. 1995. Visualizing and quantifying natural selection. Trends Ecol. Evol. 10:313–318.

Brommer, J. E., and B. Class. 2017. Phenotypic correlations capture between-individual correlations underlying behavioral syndromes. Behav. Ecol. Sociobiol. 71:8.

Brommer, J. E., K. Rattiste, and A. J. Wilson. 2008. Exploring plasticity in the wild: laying date-temperature reaction norms in the common gull *Larus canus*. Proc. R. Soc. Lond. Ser. B 275:687–693.

Bronshtein, I. N., K. A. Semendyayev, G. Musiol, and H. Mühlig. 2015. Handbook of mathematics. Springer-Verlag, Berlin Heidelberg.

Browne, W. J., R. H. McCleery, B. C. Sheldon, and R. A. Pettifor. 2007. Using cross-classified multivariate mixed response models with application to life history traits in great tits (*Parus major*). Statist. Model. 7:217–238.

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

Cleasby, I. R., S. Nakagawa, and H. Schielzeth. 2015. Quantifying the predictability of behaviour: statistical approaches for the study of between-individual variation in the within-individual variance. Methods Ecol. Evol. 6:27–37.

Davies, N. B., J. R. Krebs, and S. A. West. 2012. An introduction to behavioural ecology. Wiley-Blackwell, Oxford.

De Lisle, S. P., and E. I. Svensson. 2017. On the standardization of fitness and traits in comparative studies of phenotypic selection. Evolution 71:2313–2326.

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

Dingemanse, N. J., A. J. N. Kazem, D. Réale, and J. Wright. 2010. Behavioural reaction norms: animal personality meets individual plasticity. Trends Ecol. Evol. 25:81–89.

Dochtermann, N. A. 2011. Testing Cheverud's conjecture for behavioral correlations and behavioral syndromes. Evolution 65:1814–1820.

Fuller, W. A. 1987. Measurement error models. John Wiley and Sons, New York.

Hadfield, J. D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm *R* package. J. Statist. Softw. 33: 1–22.

Hadfield, J. D., A. J. Wilson, D. Garant, B. C. Sheldon, and L. E. B. Kruuk. 2010. The misuse of BLUP in ecology and evolution. Am. Nat. 175:116–125.

Holtmann, B., M. Lagisz, and S. Nakagawa. 2017. Metabolic rates, and not hormone levels, are a likely mediator of between-individual differences in behaviour: a meta-analysis. Funct. Ecol. 31:685–696.

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

Houslay, T. M., and A. J. Wilson. 2017. Avoiding the misuse in BLUP in behavioral ecology. Behav. Ecol. 28:948-952.

Husby, A., M. E. Visser, and L. E. B. Kruuk. 2011. Speeding up microevolution: the effects of increasing temperature on selection and genetic variance in a wild bird population. PLoS Biol. 9. https://doi.org/10.1371/journal.pbio.1000585.g002

Kingsolver, J. G., H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, A. Hoang, P. Gilbert, and P. Beerli. 2001. The strength of phenotypic selection in natural populations. Am. Nat. 157:245–261.

Kingsolver, J. G., S. E. Diamond, A. M. Siepielski, and S. M. Carlson. 2012. Synthetic analyses of phenotypic selection in natural populations: lessons, limitations and future directions. Evol. Ecol. 26:1101–1118.

Knapczyk, F. N., and J. K. Conner. 2007. Estimates of the average strength of natural selection are not inflated by sampling error or publication bias. Am. Nat. 170:501–508.

Krebs, J. R., and N. B. Davies. 1997. Behavioural ecology: an evolutionary approach. Sinauer Associates, Sunderland, MA.

Kruuk, L. E. B. 2004. Estimating genetic parameters in natural populations using the 'animal model'. Proc. R. Soc. Lond. Ser. B 359:873–890.

Lande, R. 1979. Quantitative genetics 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.

Matsumura, S., R. Arlinghaus, and U. Dieckmann. 2012. Standardizing selection strengths to study selection in the wild: a critical comparison and suggestions for the future. Bioscience 62:1039–1054.

Mitchell, D., A. M. Dujon, C. Beckmann, and P. A. Biro. 2019. Temporal autocorrelation: a neglected factor in the study of behavioral repeatability and plasticity. Behav. Ecol. 31:222–231.

Moiron, M., Y. G. Araya-Ajoy, K. J. Mathot, A. Mouchet, and N. J. Dingemanse. 2019. Functional relations between body mass and risk-taking behavior in wild great tits. Behav. Ecol. 30:617–623.

Morrissey, M. B., and J. D. Hadfield. 2012. Directional selectoin in temporally replicated studies is remarkably consistent. Evolution 66:435–442.

Morrissey, M. B., L. E. B. Kruuk, and A. J. Wilson. 2010. The danger of applying the breeder's equation in observational studies of natural populations. J. Evol. Biol. 23:2277–2288.

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

Muff, S., and L. F. Keller. 2015. Reverse attenuation in interaction terms due to covariate measurement error. Biom. J. 57:1068–1083.

Nicolaus, M., J. E. Brommer, R. Ubels, J. M. Tinbergen, and N. J. Dingemanse. 2013. Exploring patterns of variation in clutch size-density reaction norms in a wild passerine bird. J. Evol. Biol. 26: 2031–2043.

Niemelä, P. T., and N. J. Dingemanse. 2017. Individual versus pseudo-repeatability in behaviour: lessons from translocation experiments in a wild insect. J. Anim. Ecol. 86:1033–1043.

———. 2018. Meta-analysis reveals weak associations between intrinstic state and personality. Proc. R. Soc. Lond. Ser. B 1873:20172823.

Nussey, D. H., E. Postma, P. Gienapp, and M. E. Visser. 2005. Selection on heritable phenotypic plasticity in a wild bird population. Science 310:304–306.

Nussey, D. H., A. J. Wilson, and J. E. Brommer. 2007. The evolutionary ecology of individual phenotypic plasticity in wild populations. J. Evol. Biol. 20:831–844.

Phillimore, A. B., J. D. Hadfield, O. R. Jones, and R. J. Smithers. 2010. Differences in spawning date between populations of common frog reveal local adaptation. Proc. Natl. Acad. Sci. USA 107:8292–8297.

Phillips, P. C., and S. J. Arnold. 1989. Visualizing multivariate selection. Evolution 43:1209–1222.

Ponzi, E., L. F. Keller, T. Bonnet, and S. Muff. 2018. Heritability, selection, and the response to selection in the presence of phenotypic measurement error: effects, cures, and the role of repeated measurements. Evolution 72:1992–2004.

Ramakers, J. J. C., A. Culina, M. E. Visser, and P. Gienapp. 2018. Environmental coupling of heritability and selection is rare and of minor evolutionary significance in wild populations. Nat. Ecol. Evol. 2:1093–1103.

Ramakers, J. J. C., P. Gienapp, and M. E. Visser. 2019. Phenological mismatch drives selection on elevation, but not on slope, of breeding time plasticity in a wild songbird. Evolution 73:175–187.

Rausher, M. D. 1992. The measurement of selection on quantitative traits—biases due to environmental convariances between traits and fitness. Evolution 46:616–626.

Reed, T. E., P. Gienapp, and M. E. Visser. 2016. Testing for biases in selection on avian reproductive traits and partitioning direct and indirect selection using quantitative genetic models. Evolution 70:2211–2225.

Reznick, D., L. Nunney, and A. Tessier. 2000. Big houses, big cars, superfleas and the costs of reproduction. Trends Ecol. Evol. 15:421–425.

Rivkin, L. R., J. S. Santangelo, M. Alberti, M. F. J. Aronson, C. W. de Keyzer, S. E. Diamond, M. J. Fortin, L. J. Frazee, A. J. Gorton, A. P. Hendry, et al. 2019. A roadmap for urban evolutionary ecology. Evol. Appl. 12:384–398.

Robertson, B. A., J. S. Rehage, and A. Sih. 2013. Ecological novelty and the emergence of evolutionary traps. Trends Ecol. Evol. 28:552–560.

Roff, D. A. 1997. Evolutionary quantitative genetics. Chapman and Hall, New York.

Roff, D. A., and D. J. Fairbairn. 2012. A test of the hypothesis that correlational selection generates genetic correlations. Evolution 66:2953–2960.

Santangelo, J. S., L. R. Rivkin, and M. T. J. Johnson. 2018. The evolution of city life. Proc. R. Soc. B-Biol. Sci. 285:20181529.

Schlichting, C. D. 1989. Phenotypic integration and environmental change. BioScience 39:460–464.

Searle, S. R. 1961. Phenotypic, genetic and environmental correlations. Biometrics 17:474–480.

Siepielski, A. M., J. D. DiBattista, and S. M. Carlson. 2009. It's about time: the temporal dynamics of phenotypic selection in the wild. Ecol. Lett. 12:1261–1276.

Sinervo, B., and E. Svensson. 2002. Correlational selection and the evolution of genomic architecture. Heredity 89:329–338.

Snijders, T. A. B., and R. J. Bosker. 1999. Multilevel analysis—an introduction to basic and advanced multilevel modelling. Sage, London.

Stearns, S. C. 1992. The evolution of life histories. Oxford Univ. Press, New York.

Stinchcombe, J. R., M. T. Rutter, D. S. Burdick, P. Tiffin, M. D. Rausher, and R. Mauricio. 2002. Testing for environmentally induced bias in phenotypic estimates of natural selection: theory and practice. Am. Nat. 160:511–523.

Stinchcombe, J. R., A. F. Agrawal, P. A. Hohenlohe, S. J. Arnold, and M. W. Blows. 2008. Estimating nonlinear selection gradients using quadratic regression coefficients: Double or nothing? Evolution 62:2435–2440.

Thomson, C. E., F. Bayer, N. Crouch, S. Farrell, E. Heap, E. Mittell, M. Zurita-Cassinello, and J. D. Hadfield. 2017. Selection on parental performance opposes selection for larger body mass in a wild population of blue tits. Evolution 71:716–732.

van de Pol, M., and J. Wright. 2009. A simple method for distinguishing within- versus between-subject effects using mixed models. Anim. Behav. 77:753–758.

van Noordwijk, A. J., and G. de Jong. 1986. Acquisition and allocation of resources—their influence on variation in life-history tactics. Am. Nat. 128:137–142.

Videlier, M., V. Careau, A. J. Wilson, and H. D. Rundle. 2020. Quantifying selection on standard metabolic rate and body mass in *Drosophila melanogaster*. Evolution 75:130–140.

Westneat, D. F., Y. G. Araya-Ajoy, H. Allegue, B. Class, N. J. Dingemanse, N. A. Dochtermann, L. Z. Garamszegi, J. G. A. Martin, S. Nakagawa, D. Reale, et al. 2020. Collision between biological process and statistical analysis revealed by mean-centering. J. Anim. Ecol. 89:2813–2824.

Westneat, D. F., and C. W. Fox. 2010. Evolutionary behavioural ecology. Oxford Univ. Press, New York.

Westneat, D. F., M. I. Hatch, D. P. Wetzel, and A. L. Ensminger. 2011. Individual variation in parental care reaction norms: integration of personality and plasticity. Am. Nat. 178:652–667.

Associate Editor: Dr. Joseph Tobias
Handling Editor: Dr. David Hall

Supporting Information

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

Table S1. Studies publishing estimates of linear and nonlinear selection in *Evolution* from 2010 to 2019, with species, trait studied, category of trait 44 (MO = morphological, BEH = behavioral, LH = life history, PHY = physiological, PER = performance), fitness measure (L = lifetime, typically survival; E = one measure of an episode of fitness; E2 = at least two measures of episodic fitness), number of measures taken, whether the mean was used if more than one measure (or if ≥2 traits were combined with PCA), whether repeatability was mentioned and its magnitude if known, type of selection measured (D = directional, Q = quadratic, C = correlational), whether multivariate models were used, if traits were mean-centered before analysis (? = either authors did not say or simply stated they "standardized" without defining; residuals and PCA were counted as mean-centered) and if among-trait correlations were provided in cases of nonlinear selection. Entries left blank if non-applicable.

Figure S3. Illustration of a parabolic relationship between trait (t) on absolute fitness (W), where the dotted line represents the population-mean trait value, the star represents the optimal trait value; (**a**) the orange dot represents the tangent line where the trait value has the value zero. (**b**) the blue dot represents the tangent line at the population-mean trait value.

Table S7. Estimates of accuracy and precision in linear (β1), quadratic (γ11), and correlational (γ12) selection gradients derived from regression models fitting one observed trait value or a mean of three observed trait values, multivariate mixed-effects models, and errors-in-variables models.

Supporting Information