Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution doi: 10.1111/j.2041-210X.2011.00160.x

Quantifying individual variation in reaction norms: how study design affects the accuracy, precision and power of random regression models

Martijn van de Pol*

Evolution, Ecology & Genetics, Research School of Biology, Australian National University, 0200 ACT, Canberra, Australia

Summary

1. Quantifying individual heterogeneity in plasticity is becoming common in studies of evolutionary ecology, climate change ecology and animal personality. Individual variation in reaction norms is typically quantified using random effects in a mixed modelling framework. However, little is known about what sampling effort and design provide sufficient accuracy, precision and power.

2. I developed '*odprism*', an easy-to-use software package for the statistical language R, which can be used to investigate the accuracy, precision and power of random regression models for various types of data structures. Moreover, I conducted simulations to derive rules-of-thumb for four design decisions that biologists often face.

3. First, I investigated the trade-off between sampling many individuals a few times versus sampling few individuals often. Generally, at least 40 individuals should be sampled with a total sample size of at least 1000 to obtain accurate and precise estimates of individual variation in elevation and slopes of linear reaction norms and their correlation. Contrasting a previous recommendation, it is worthwhile to bias the ratio of number of individuals over replicates towards sampling more individuals.

4. Second, I considered how the range of environmental conditions over which individuals are sampled affects the optimal sampling strategy. I show that when all individuals experience the same conditions during a sampling event, sampling each individual only twice should be strictly avoided.

5. Third, I examined the case where the number of replicates per individual is constrained by their lifespan, as is common when sampling annual traits in the wild. I show that for a given sampling effort, it is much easier to detect individual variation in reaction norms for long-lived than for short-lived species.

6. Fourth, I investigated the performance of random regression models when studying traits under selection. Reassuringly, directional viability selection barely caused any bias in estimates of variance components.

7. Random regression models are inherently data hungry, and reviewing the literature shows that particularly behavioural studies have low sampling effort. Therefore, the software and rules-of-thumbs I identified for designing reaction-norm studies should help researchers make more informed choices, which likely improve the reliability and interpretation of plasticity studies.

Key-words: annual traits, bias, I × E, life-history, lifespan, mixed model, phenotypic plasticity, random slopes, statistics, viability selection

Introduction

Phenotypic plasticity, the change in the phenotype of an organism in response to environmental change, is an important source of trait variation (Pigliucci 2001). Interest in individual variation in phenotypic plasticity is growing apace, because it is increasingly recognized that such variation is important from evolutionary, ecological as well as behavioural perspectives. Natural selection not only acts on variation in the mean trait value, but can also act on variation in levels of plasticity; if individual variation in plasticity has a heritable basis, such

*Correspondence author. E-mail: m.van.de.pol@myscience.eu
Correspondence site: http://www.respond2articles.com/MEE/

selection directly impacts the course of microevolution (Nussey *et al.* 2005; Nussey, Wilson & Brommer 2007). Similarly, the degree of plasticity is thought to be a key determinant of the ability of populations to respond to climate change, which ultimately determines whether numbers decline or not (Visser 2008; Chevin, Lande & Mace 2010; Reed *et al.* 2010). Individual variation in plasticity also influences the amount of individual heterogeneity in vital rates and thereby directly affects population dynamics and extinction risk (Vindenes, Engen & Saether 2008). Finally, in the field of animal personality, individual variation in behavioural personality and behavioural plasticity are considered complementary aspects of the individual phenotype (Sih *et al.* 2004; Dingemanse *et al.* 2010).

Phenotypic plasticity in labile continuous traits is typically conceptualized in terms of reaction norms, functions that relate individual phenotypes to an environmental variable (de Jong 1990; Pigliucci 2001). In its simplest form, the linear reaction-norm approach is characterized by two components: elevation and slope. If the environmental covariate of interest is mean-centred, the elevation component reflects the expected trait value in the average environment. The slope component describes the change in phenotype across an environmental gradient and is therefore a measure of phenotypic plasticity.

Individual variation in reaction norms is usually quantified using a specific type of mixed model, the random regression model (Henderson 1982), which estimates the amount of among-individual variation in both elevations and slopes of reaction norms (see Methods). Significant variance in slopes among individuals can be interpreted as an individual by environment interaction ('I × E'; Nussey, Wilson & Brommer 2007), which may provide the basis for further investigations into whether such heterogeneity is caused by genetic or permanent environmental effects. Several recent papers have advocated the use of random regression models in various fields of biology (e.g. van de Pol & Verhulst 2006; Nussey, Wilson & Brommer 2007; van de Pol & Wright 2009; Dingemanse *et al.* 2010), and they are being used with increasing frequency (see overview in Appendix S1). Even if one is not interested in individual variation in plasticity *per se*, it can still be important to model such heterogeneity to obtain reliable estimates of population levels of plasticity (Schielzeth & Forstmeier 2009).

Despite the popularity of random regression models to quantify individual variation in reaction norms, very little is known about the data requirements for such models and how this affects the accuracy, precision and power of reaction-norm analysis (Martin *et al.* 2011). When designing a study, such knowledge is important, as systematic bias in parameters can result in incorrect interpretation of the amount of individual variation in phenotypic plasticity. Low precision increases the likelihood of extreme outcomes for individual case studies, which also gives the impression that there is little pattern across studies. And, lack of adequate statistical power results in frequent null results, which are difficult to interpret and also frustrate meta-analyses because of publication bias towards statistically significant results.

Unfortunately, determining the optimal study design is typically complicated for mixed models, and unambiguous conclusions can be difficult to make (Scherbaum & Ferreter 2009). The difficulty arises from the complexity of these models, in which (i) both the number of individuals and the number of replicates per individual sampled can vary, including unbalanced designs where some individuals are sampled more often than others, (ii) there are many fixed (e.g. mean slope) and random parameters (e.g. variance in slopes among individuals) of interest, and maximizing each of their precision may exert different demands on the optimal sampling design and (iii) the optimal sampling design may depend on patterns in the data, such as the degree with which traits vary among and within individuals, which in turn is contingent on the biological context. Moreover, accuracy and precision can also depend on the algorithm used to estimate parameters (Maas & Hox 2004b).

Scherbaum & Ferreter (2009) and Hox (2010) provide comprehensive overviews of the statistical literature on sample size and power analysis in mixed models. A rule-of-thumb appears to be that random parameters can typically be estimated with lower precision than fixed parameters (Hox 2010). Also, a large sample of individuals is typically more important than a large number of replicates per individual (Maas & Hox 2004b; Hox 2010), although when interested in parameters that vary within individuals, it can – in specific circumstances – be important to focus sampling effort on replicates (Snijders 2005). Finally, low sample size can result in biased estimates of variance components and their standard errors as well as low power (Verbeek 2000). Notwithstanding, recommendations for minimum sample sizes vary widely (ranging from 10 to 100 individuals; Snijders & Bosker 1999; Maas & Hox 2004b), while Martin *et al.* (2011) suggested that power to statistically detect individual heterogeneity in slopes is highest if the ratio of number of individuals over replicates is a half.

The study by Martin *et al.* (2011) may be particularly relevant, as it specifically focussed on the power of random regression models while addressing issues that are inherent to data collection in an ecological and reaction-norm context, such as censored unbalanced data and the type of environmental variability. However, Martin *et al.* (2011) did not investigate how sampling strategy affects the accuracy and precision of variance component estimates, while quantitative biologists may be at least as interested in using accuracy and precision as criteria for study design than statistical power alone.

Here I develop and introduce an easy-to-use flexible set of functions (library '*odprism*'; van de Pol 2011) for the freely available statistical language R (R Development Core Team 2010), which can be used to investigate the accuracy, precision and power of random regression models for various types of data structures commonly encountered in biology. Building on the work by Martin *et al.* (2011), I subsequently present results on data simulations which are used to derive rules-of-thumb for four types of study design decisions that biologists often face: (i) the trade-off between sampling many individuals a few times versus sampling few individuals often, (ii) the range of environmental conditions over which individuals are sampled, (iii) how the longevity of study species affects study design when the number of replicates per individual is constrained by their lifespan, as is common when sampling annual traits in the

Table 1. Explanation of abbreviations and definition of key concepts used in the paper

Term	Meaning and definition
Estimated parameters	
V_E	Among-individual variance in elevation (intercept)
V_S	Among-individual variance in slope
$r_{E,S}$	Correlation among individuals' elevation and slope
V_R	Within-individual (residual) variance
Sampling variables	
I	Number of individuals sampled
J	Number of replicates sampled per individual
P	Population size sampled in each year
T	Number of years population is sampled
N	Total sample size (note that $N = I \times J$ and $N = P \times T$)
Q	Annual adult survival probability
Model performance	
Accuracy & bias	(In)accuracy is used as a qualitative term describing the (dis)agreement between model estimates and the 'true' value, while bias is used as a quantitative term, operationalized as the difference between the median model estimate of the 5000 simulated data sets and the value used to generate the data sets, e.g. the relative bias (%) in V_E equals $\left\lvert \hat{V}_E - V_E \right\rvert / V_E \times 100$
Precision	(Im)precision describes the degree to which different simulations give (dis)similar results, quantified by the width of the distribution of parameter estimates from 5000 simulated data sets, i.e. here the difference between the 75% and 25% percentile
Statistical power	Proportion of mixed models applied to 5000 simulated data sets that correctly rejected the false null hypothesis of no effect (i.e. $P < 0{\cdot}05$)
Model convergence	Proportion of mixed models applied to the 5000 simulated data sets that did not convergence or reported convergence problems

wild and (iv) the accuracy of random regression models when studying traits under viability selection, which may be poor when survivors with specific trait values are overrepresented.

Methods

In Table 1, I provide an overview of the abbreviations and key definitions used throughout the paper. For reasons of simplicity, I focus on continuous labile traits that follow a normal distribution and exhibit linear reaction norms. Moreover, given the interest in plasticity, I only discuss individual's phenotypic responses to environmental variables; however, this framework is equally relevant for intrinsic variables (e.g. age, physiological state). Similarly, I consider data to be grouped within individuals, but the framework also applies to data structured by other grouping variables, for example a social group, population or genotype.

RANDOM REGRESSION MODELS

Random regression models are mixed models that contain both a random intercept and random slope term, and routinely estimate three key parameters: (i) the variance among individuals in intercepts (elevations), (ii) the variance among individuals in slopes and (iii) the covariance (or correlation) between individual's intercepts and slopes (e.g. Snijders & Bosker 1999). The model can be described by:

$$Y_{ij} = \beta_0 + u_{0i} + (\beta_1 + u_{1i})X_{ij} + e_{0ij} \qquad \text{eqn 1a}$$

where subscripts i and j refer to the structuring of the data, with Y_{ij} being the value of the dependent variable and X_{ij} to the value of the environment for measurement j from individual i. Furthermore, the intercept of the regression equation β_0 can be interpreted as the population mean when the environmental variable is mean-centred ($\bar{X} = 0$). The population slope (i.e. plasticity) of the dependency of Y on X is given by β_1. The error term e_{0ij}, the random intercept u_{0i} and the random slope term u_{1i} are typically assumed to be drawn from (multivariate) normal distributions with expectations 0 and (co)variances such that:

$$e_{0ij} \sim \text{Normal}(0, \sigma^2_{e_{0ij}}) \text{ and } \begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim \text{MVNormal}(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma^2_{u_{0i}} & \\ \sigma_{u_{0i},u_{1i}} & \sigma^2_{u_{1i}} \end{bmatrix} \qquad \text{eqn 1b}$$

The term $\sigma^2_{e_{0ij}}$ is used to estimate the within-individual residual variance (V_R), the term $\sigma^2_{u_{0i}}$ is used to estimate the among-individual variance in elevation (V_E), the term $\sigma^2_{u_{1i}}$ is used to estimate the among-individual variance in slopes (V_S), and the covariance term $\sigma_{u_{0i},u_{1i}}$ is used to estimate the correlation between the intercepts and slopes of individuals ($r_{E,S}$). I prefer to report the correlation instead of the covariance, because a correlation lies between −1 and 1, allowing one to conclude that a model with a correlation close to those extremes is weakly identifiable.

GENERAL APPROACH OF SIMULATIONS

The general approach in all four scenarios described below was to generate data sets of total sample size N consisting of I individuals and J replicates per individual sampled. For each data set, I generated X-values using a standard normal random variable [$X \sim$ Normal(0,1)]. By inserting these X-values into eqns 1a & 1b, I generated Y-values, whereby all fixed effects were set to 0 ($\beta_0 = \beta_1 = 0$) and whereby error and random terms were generated by drawing from (multivariate) normal random variables with expectations 0 and (co)variances equal to V_R,V_E, V_S, $r_{E,S}$. This process was repeated to generate 5000 different data sets for each combination of sampling design and values of V_R,V_E, V_S, $r_{E,S}$.

In all scenarios, I standardized the expected total variance at $X = 0$ to one by setting $V_E + V_R = 1$, which implies that the intraclass

correlation coefficient τ at $X = 0$ – a measure of repeatability – has the same value as V_E [since $\tau_{X=0} = V_E/(V_E + V_R)$]. Inspired by repeatabilities commonly observed in biological studies (Martin *et al.* 2011), I investigated the parameter conditions $\tau_{X=0} = V_E = 0.2$ or $\tau_{X=0} = V_E = 0.4$ and explored the conditions $V_S = 0.1$ or $V_S = 0.2$ and $r_{E,S} = 0.25$ or $r_{E,S} = 0.5$. The values 2–10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150, 200 were sampled for both I and J, and bilinear interpolation between these integer values was used to determine isoclines of similar accuracy, precision and power.

SCENARIO 1: SAMPLING MANY INDIVIDUALS OR MANY REPLICATES?

In scenario 1, I explore the trade-off between sampling many individuals a few times versus sampling few individuals often. Given that one can only sample a limited number of times, what is a reasonable number of individuals and how often should one sample each individual?

SCENARIO 2: TYPE OF ENVIRONMENTAL VARIABILITY

In scenario 1, all individuals experienced different environmental conditions when sampled at a given occasion. Such a situation reflects, for example, the case where the environmental variable is food abundance, and food abundance varies between territories where individuals are sampled (e.g. at sampling occasions 1 and 2, individual A experienced food abundance 1·2 and 2·3, while individual B experienced food abundance 3·5 and 0·9). However, in many situations, the environmental conditions might be the same for all individuals on a given occasion. For example, when studying the effect of spring temperature on the timing of egg-laying, all birds sampled in a specific breeding season have experienced the same spring temperature (e.g. at sampling occasions 1 and 2, individual A experienced temperatures 17·2 and 19·1, and individual B also experienced temperatures 17·2 and 19·1). In experimental studies, one typically also exposes each individual to the same set of environmental conditions. In scenario 2, I therefore explore the influence of the type of environmental variability, by using sampling strategies where all individuals have either a different or the same value of environmental variable X on a given sampling occasion.

SCENARIO 3: CONSTRAINTS IMPOSED BY THE STUDY SPECIES

In scenario 3, I address a typical problem encountered when sampling annual traits in the wild. When sampling traits that are only measured or expressed once a year (e.g. body mass at the start of the reproductive season, timing of flowering), one is not completely free to choose the number of individuals and replicates sampled. Some individuals may live longer than others and when following a population of individuals over time, the number of replicates per individual will ultimately be constrained by an individual's lifespan. Censoring after the last year of study is another mechanism constraining the number of replicates per individual sampled. Even though mortality precludes complete control over the number of individuals and replicates sampled in a long-term population study, their numbers can be predicted. Assuming that death individuals are replaced (i.e. constant population size), the expected number of individuals $E(I)$ that can be sampled depends on the annual survival Q, monitored population size P and number of years monitored T:

$$E(I) = P + P(1 - Q)(T - 1) \quad \text{eqn 2a}$$

The expected number of replicates per individuals $E(J)$ depends on a species annual survival and the study period but is independent of the study's population size:

$$E(J) = \frac{PT}{E(I)} = \frac{T}{1 + (1 - Q)(T - 1)} \quad \text{eqn 2b}$$

Another key consequence of sampling annual traits is that – in contrast to the balanced data sets in scenarios 1 and 2 – variation in lifespan results in unbalanced data sets, with some individuals measured only once or twice, while others measured up to, for example, ten times. Small differences in annual survival between species can result in large differences in life expectancy (Fig. 1a), which profoundly influences the distribution of number of replicates per individual obtainable from species of varying longevity (Fig. 1b). In a long-term population study, one cannot expect to be able to measure most individuals more than once in short-lived species (annual survival $\leq 0{\cdot}5$), while in long-lived species (survival $\geq 0{\cdot}8$), one can expect to be able to measure the majority of individuals at least thrice (Fig. 1b). This begs the question to what extent species longevity – and thereby choice of study species – affects the ability to reliably quantify individual heterogeneity in reaction norms.

Therefore, I generated data sets with population size P that were followed over a number of years T, in which I varied the annual survival probability Q between 0·3 and 1·0, whereby Q was assumed to be the same for all individuals and invariable with age. When an individual died, it was replaced by a new individual in the data simulation process (i.e. no inheritance of trait values). All individuals were given the same value of environmental variable X on a given sampling occasion. The values 2–10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150, 200 were sampled for both P and T, and bilinear interpolation between these integer values was used to determine isoclines of similar accuracy, precision and power.

SCENARIO 4: TRAITS UNDER VIABILITY SELECTION

In scenario 4, I consider the question to what extent viability selection on annual traits affects the accuracy of estimates of model parameters. Awareness is growing that viability selection may cause bias in estimating various evolutionary parameters of interest (Vaupel, Manton & Stallard 1979; van de Pol & Verhulst 2006; Hadfield 2008). When survival depends on a trait value, viability selection causes the population of survivors to be a nonrandom subset of the population, which Martin *et al.* (2011) hypothesized could also bias reaction-norm parameters.

To investigate the effects of viability selection, I generated data sets similar as in scenario 3 but made individuals' survival probability Q from year t to $t + 1$ dependent on the trait value Y in year t. I considered viability selection generated by the logistic function:

$$Q_{it} = 1/(1 + \exp(-(\gamma_0 + \gamma_1 Y_{it}))) \quad \text{eqn 3}$$

and varied γ_0 between −0·8 and 2·2 to explore the impact of variation in mean annual survival (i.e. species longevity), and varied γ_1 between 0 and 10 to investigate the impact of the strength of directional selection. To maximize chances of detecting any bias, data sets considered large populations (i.e. $P = 1000$, $T = 10$).

PERFORMANCE OF RANDOM REGRESSION MODELS

I assessed four measures of model performance: accuracy\bias, precision, statistical power and model convergence; Table 1 describes how

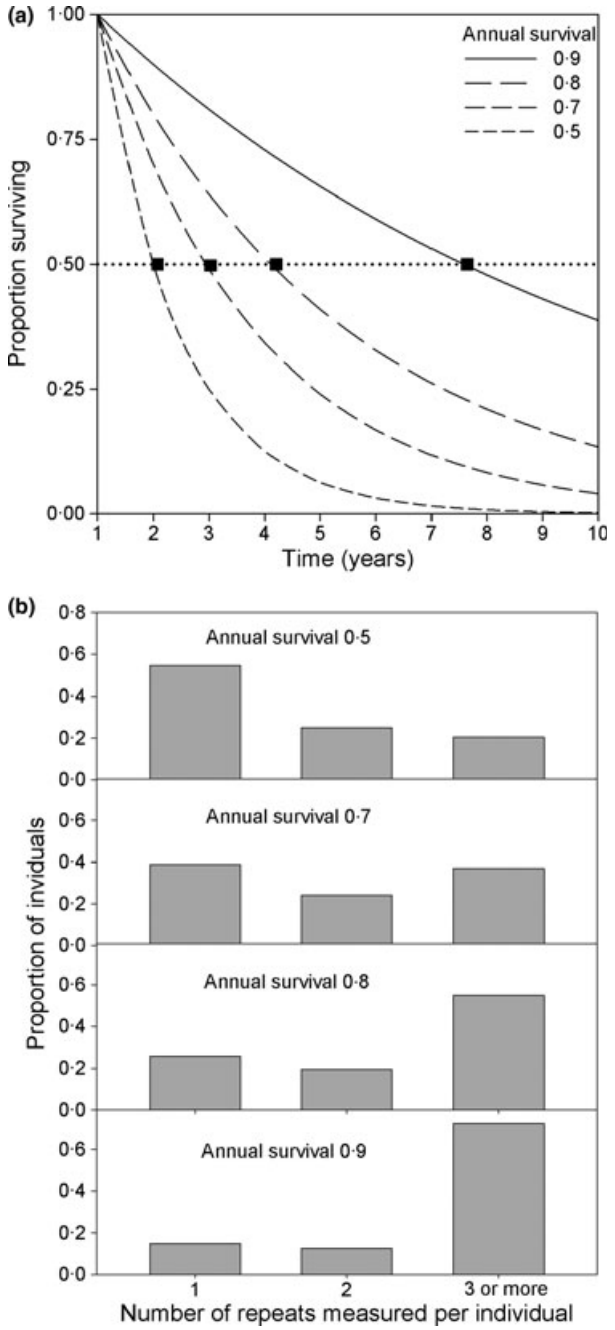

Fig. 1. (a) Survival curves of hypothetical species with different annual adult survival probabilities ranging from short-lived (0·5) to long-lived (0·9); the adult life expectancy for each of the species is depicted by the black square symbols. (b) The expected frequency distribution of the number of repeat observations per individuals if a population of each of the hypothetical species would be followed over a long period.

they are defined and operationalized. I did not explore the accuracy of the standard errors of random parameters, because they are likely to be asymmetric and thus difficult to interpret. I assessed statistical power to detect among-individual variation in slopes by comparing the likelihood of the model described in eqns 1a & 1b with the likelihood of a model where the terms used to estimate V_S and $r_{E,S}$ were both constrained to 0 (i.e. $\sigma^2_{u_{1i}} = \sigma_{u_{0i},u_{1i}} = 0$). Individual variation in elevation was tested for by comparing the likelihood of a model that contained only a random intercept with the likelihood of a null model where the term used to estimate V_E was constrained to 0 (i.e. $\sigma^2_{u_{0i}} = 0$). To obtain *P*-values, differences in $-2 \times$ log-likelihood were compared to a chi-square-distribution with the degrees of freedom equal to the number of constrained parameters (likelihood ratio test; Pinheiro & Bates 2000). Alternative, potentially better methods exist for statistical inference (Visscher 2006; Scheipl, Greven & Kuchenhoff 2008) and model selection (e.g. information theoretic approaches); however, I focus on the widely used likelihood ratio test that also allows for comparison with results from Martin *et al.* (2011).

EASY-TO-USE FUNCTIONS IN R: PACKAGE 'ODPRISM'

Existing software for assessing the optimal design of mixed models, such as '*PINT*' (Snijders & Bosker 1993), '*Optimal Design*' (Raudenbush *et al.* 2011) and '*pamm*' (Martin *et al.* 2011), does not allow for most of the four scenarios to be investigated. Therefore, I developed the package '*odprism*' (optimal design and performance of random intercept and slope models; van de Pol 2011), which contains a set of easy-to-use flexible functions written in the statistical language R that can (i) simulate all the types of data sets from scenarios 1–4, (ii) run random regression models on these data sets and (iii) quantify model performance. The functions in *odprism* use the '*lmer( )*' function from the R-package '*lme4*' (Bates, Maechler & Dai 2008) to estimate the model parameters based on a restricted maximum likelihood (REML) approach in which estimated variance–covariance matrices are constrained to be positive-definite. The source code of *odprism* and its manual with example code are available on the Comprehensive R Archive Network (http://cran.r-project.org/web/packages/odprism/).

Results

In all scenarios, convergence problems were rare (< 1%) as long as total sample size exceeded $N = 20$; therefore, I henceforth only present results on accuracy, precision and power. Also, I focus on the results for parameter conditions $V_E = 0{\cdot}2$, $V_S = 0{\cdot}1$ and $r_{E,S} = 0{\cdot}5$, while results for other parameter conditions ($V_E = 0{\cdot}4$, $V_S = 0{\cdot}2$ or $r_{E,S} = 0{\cdot}25$) are presented in Appendices S1–S6.

SCENARIO 1: SAMPLING MANY INDIVIDUALS OR MANY REPLICATES?

For a given sampling effort, there are many possible combinations to sample I individuals J times (see lower and upper *x*-axes in Fig. 2). Reproducing the results from the study by Martin *et al.* (2011), I also found that (i) power to detect individual variation in elevation is typically higher than power to detect individual variation in slopes (Fig. 2a–c), (ii) power is always high if $N \geq 1000$ (Fig. 2c) and (iii) power appears to be maximized at a ratio of I/J of 0·5 (at five individuals sampled 10 times each if $N = 50$, Fig. 2a; at 10 individuals sampled 20 times each if $N = 200$, Fig. 2b).

However, a sampling strategy that generates high statistical power to detect individual variation in intercept and slopes does not necessarily ensure that the accuracy and precision of these terms are also high. Specifically, there always was an

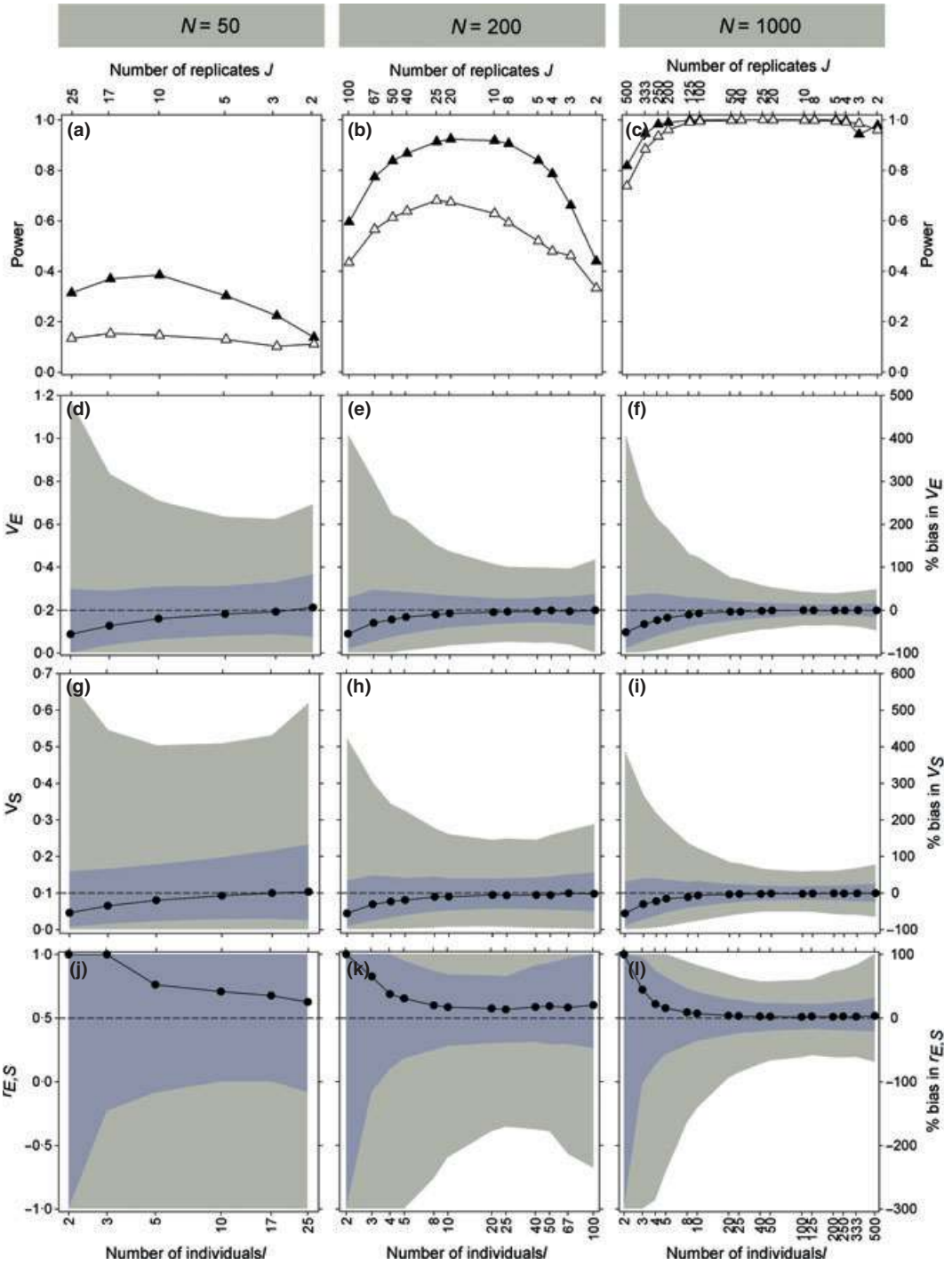

Fig. 2. (a–c) Power of detecting individual heterogeneity in elevation (black triangles) and slopes (white triangles) and accuracy and precision of V_E (d–f), of V_S (g–i) and of $r_{E,S}$ (j–l) for different total sample size N (left, middle or right panels) made up of different combinations of number of individuals I (lower x-axis) and replicates per individual J (upper x-axis). In (d–l), circles represent the median estimates of 5000 simulated data sets, while the dark and light grey areas depict, respectively, the 25–75% and 2·5–97·5% distribution of parameter estimates. The horizontal dashed lines depict the values used to generate the data ($V_E = 0{\cdot}2$, $V_S = 0{\cdot}1$, $r_{E,S} = 0{\cdot}5$). Note that the right y-axes depict the relative deviation from the values used to generate the data.

upward bias for $r_{E,S}$ at low total sample sizes ($N \leq 200$; Fig. 2j–k) and for $I \leq 10$ at large sample size (Fig. 2l). Moreover, independent of total sample size, both V_E (Fig. 2d–f) and V_S (Fig. 2g–i) were biased downward when $I \leq 10$. Finally, the imprecision of (co)variance components – described by the width of the distribution of parameter estimates shown in grey areas in Fig. 2d–l – was also minimized at a much higher number of individuals ($I \geq 10$ if $N = 50$; $I \geq 20$ if $N = 200$; $I \geq 40$ if $N = 1000$) than required for maximizing power.

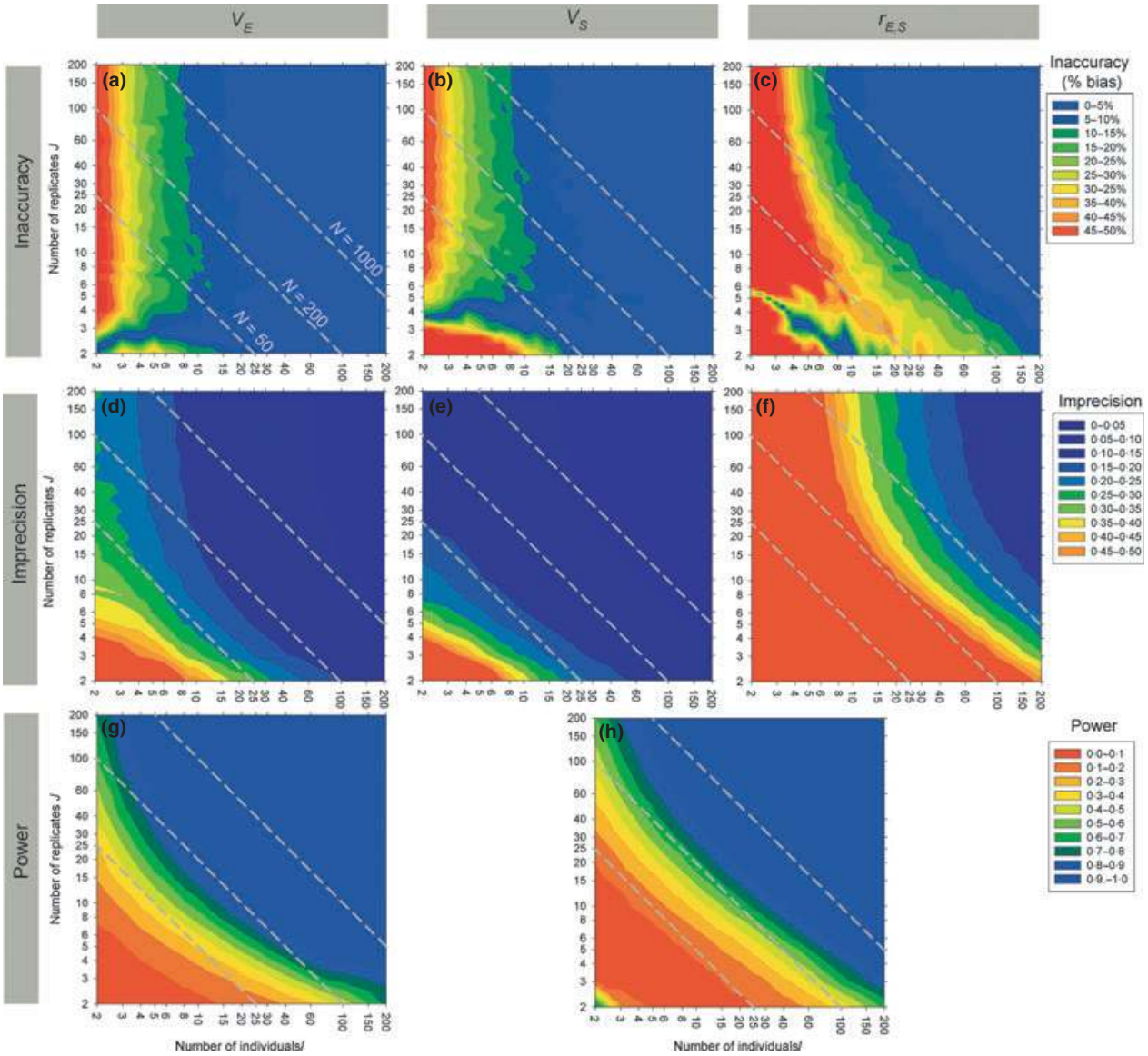

Fig. 3. Inaccuracy (relative bias; a–c) and imprecision (d–f) of estimates of random regression variance components V_E, V_S and $r_{E,S}$ and the statistical power to detect individual heterogeneity in elevation (g) and slopes (h; a combined test of both V_S and $r_{E,S}$) as a function of the number of individuals I and replicates J sampled. The definitions of accuracy, precision and power are given in Table 1. Different colours depict areas between isoclines of similar levels of inaccuracy, imprecision and power (see legends); isoclines were determined by bilinear interpolation between the sampled integer values of I and J. The values used to generate the data were $V_E = 0{\cdot}2$, $V_S = 0{\cdot}1$ & $r_{E,S} = 0{\cdot}5$.

A more comprehensive pattern about optimal sampling strategies emerges when considering how accuracy, precision and power change when one allows I and J to vary independently. To avoid a substantial systematic bias (> 10%) in estimating V_E and V_S, I should be at least 10–20 (independent of J; Fig. 3a,b), while to avoid substantial bias in $r_{E,S}$, one should sample at least 10 individuals and total sample effort should be $N \geq 300$ (e.g. sample 10 individuals 30 times or 150 individuals two times; Fig. 3c). Obtaining high precision is hardest for the correlation parameter $r_{E,S}$ and to obtain an imprecision of 0·2 (meaning that in the absence of bias, there is a 50% chance that the estimate is within 0·1 of the 'true' value of e.g. $r_{E,S} = 0{\cdot}5$), one would need to sample at least 40 individuals 40 times, or more efficiently in terms of minimizing N: sample 100 individuals 10 times or 200 individuals five times (i.e. $I \geq 40$, $N \geq 1000$ & $I/J \geq 1$ Fig. 3f). Finally, to obtain power of at least 0·8 for both the test of individual variation in elevation and slopes, total sample size should be at least 300 and the ratio of I/J should not be too skewed (i.e. roughly between 0·2 and 3; Fig. 3g,h). Thus, when interested in optimizing accuracy, precision as well as statistical power, obtaining high precision of $r_{E,S}$ puts the strongest demand on sampling design, and if one selects a sampling design that optimizes the precision of $r_{E,S}$, this should also provide precise and unbiased estimates of other covariance components as well as high power.

Most of the above patterns (presented for $V_E = 0{\cdot}2$, $V_S = 0{\cdot}1$ & $r_{E,S} = 0{\cdot}5$) seem to be fairly robust to changing

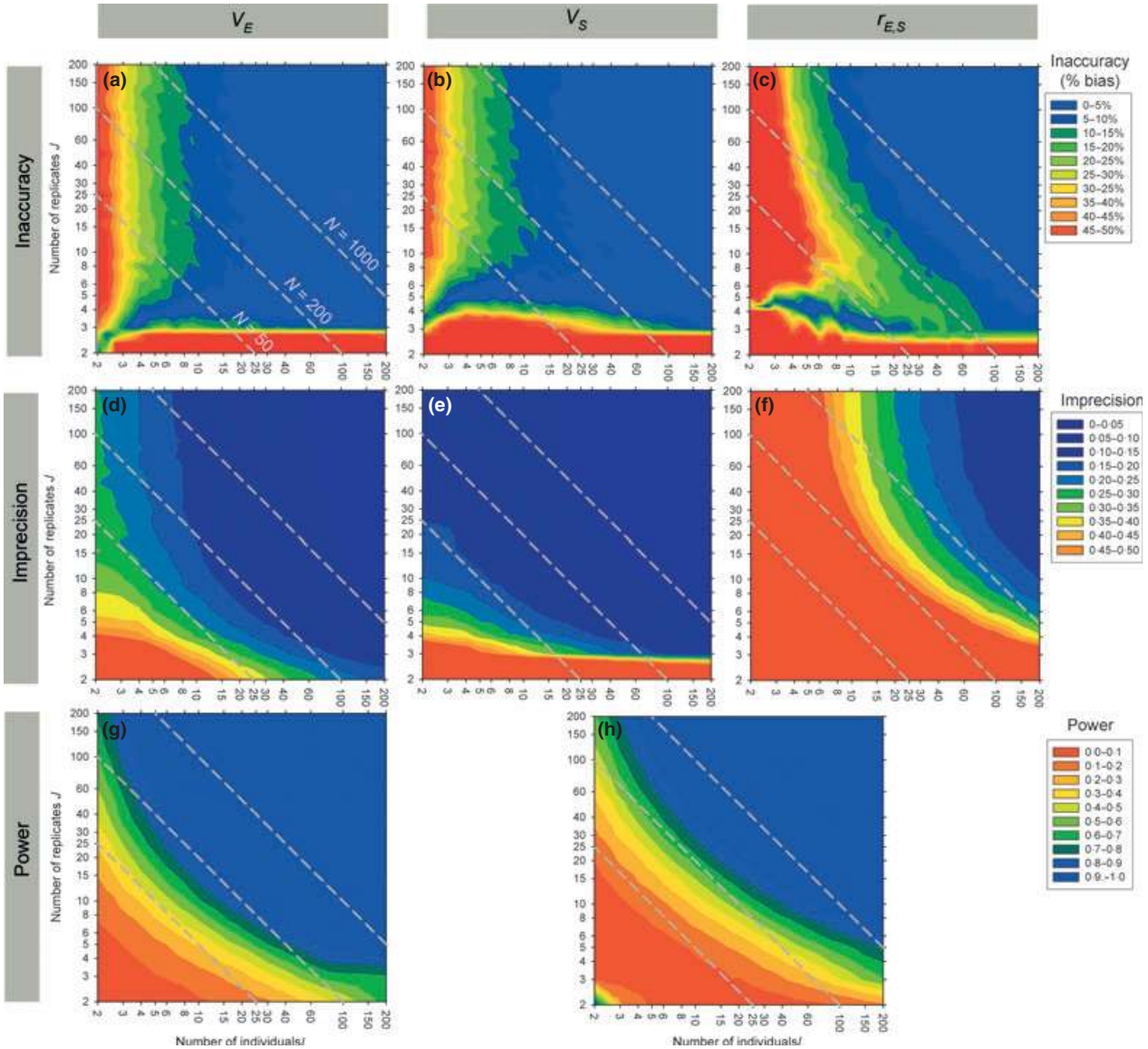

Fig. 4. Same as Fig. 3, with the difference that all individuals experienced the same environmental conditions at a given sampling occasion.

the parameter conditions used to generate the data (i.e. when doubling V_E or V_S or halving $r_{E,S}$, see Appendix S2), although this search of the parameter space is not exhaustive. The main difference when dealing with a situation of higher values of V_E, V_S and $r_{E,S}$ seems to be that for a given sampling design imprecision increases slightly, but also the statistical power (Figs B1–B3 in Appendix S2 vs. Fig. 3). Although dealing with higher values of V_E, V_S and $r_{E,S}$ did not affect the accuracy of these parameters much (Figs B1–B3 in Appendix S2 vs. Fig. 3), it should be realized that I express accuracy in terms of relative bias, implying that doubling or halving these parameters means that the absolute bias also roughly doubled or halved.

In summary, when choosing a sampling design to obtain unbiased ($< 10\%$) and precise (50% of estimates within 0·1 of the 'true' value) estimates of all variance components, while also having high statistical power ($> 0{\cdot}8$), the general rule-of-thumb appears to be to sample at least 40 individuals and have a total sample size of at least 1000. If these two conditions are fulfilled, the choice between sampling many individuals or replicates does not matter too much as long as one biases sampling towards more individuals ($I/J \geq 1$).

SCENARIO 2: TYPE OF ENVIRONMENTAL VARIABILITY

When comparing a situation where all individuals experienced different environmental conditions when sampled at a given occasion (Fig. 3) with a situation where all individuals experienced the same environmental conditions at a given sampling occasion (Fig. 4), there is one major incongruity. Sampling all individuals only twice should be strictly avoided when all individuals experienced the same environmental conditions at a given occasion. In such a situation, it becomes impossible to reliably decompose variance components, resulting in extreme bias in all variance component estimates (Fig. 4a–c). Variance

components are also quite biased and imprecise when sampling all individuals three or four times, but this problem can be overcome by sampling many individuals (Fig. 4a–f). The above problem disappears when sampling individuals five times or more (compare Fig. 4 with Fig. 3). Similar results were obtained for other values of V_E, V_S and $r_{E,S}$ (see Appendix S3). Thus, as long as one avoids sampling individuals only twice, thrice or four times, the demands on sampling design are almost identical for a situation where all individuals experienced either the same or different environmental conditions at a given sampling occasion.

It is important to realize that the comparison between a situation where all individuals experienced either different (Fig. 3) or the same (Fig. 4) environmental conditions at a given sampling occasion represent the two extreme cases. For example, environmental variables that differ between individuals at a given occasion, such as food abundance, can still be correlated between individuals at a given sampling occasion (e.g. owing to 'year' effects or spatial autocorrelation in the environment). Furthermore, when environmental variables are the same for all individuals at a given occasion, this does not preclude individual variation in the environment that individuals experienced: when one would sample on three occasions, but would only sample each individual on two of the three occasions, some individuals could experience conditions A&B, others B&C and even others A&C. In this latter scenario, it is possible to reliably decompose variance components and obtain unbiased estimates, despite the fact that individuals are only sampled twice (Fig. D1 in Appendix S4).

SCENARIO 3: CONSTRAINTS IMPOSED BY THE STUDY SPECIES

When sampling traits that are measured or expressed annually, the number of replicates per individual ultimately depends on the lifespan of individuals, causing data sets to be unbalanced. A direct consequence of this constraint is that for a given sampling design (population of size P, sampled for T years), the longevity of a study species affects the estimation of individual variation in elevation and slopes in a long-term population study (Fig. 5). In short-lived species, both statistical power (Fig. 5a,b) and precision (Fig. 5c–h) are lower than in long-lived species; similar results were obtained for other parameter values of V_E, V_S and $r_{E,S}$ (see Appendix S5). This positive association between species longevity and power and precision is likely to be a direct result from the fact that studying short-lived species precludes one from measuring most individuals more than once (Fig. 1b), while individuals measured only once do not contribute directly to the estimation of individual variation in slopes. When aiming to obtain unbiased ($< 10\%$) and precise (50% of estimates within 0·1 of the 'true' value) estimates of all (co)variance components with high statistical power ($> 0{\cdot}8$), the minimum sampling requirement for a long-lived species with annual survival of 0·9 is to have a total sample size of 1000, while for short-lived species with annual survival of 0·5, one requires a total sample size of at least 1500 (Appendix S6).

SCENARIO 4: TRAITS UNDER VIABILITY SELECTION

There was no strong support for the hypothesis that directional viability selection on traits affects the accuracy of estimates of the variance components of reaction-norm parameters. None of the (co)variance components was biased by more than 2% in both short- and long-lived species, even when directional selection was very strong (i.e. γ_0 was varied between $-0{\cdot}8$ and 2·2, γ_1 between 0 and 10; using $V_E = 0{\cdot}2$ or 0·4, $V_S = 0{\cdot}1$ or 0·2 $r_{E,S} = 0{\cdot}25$ or 0·5, $P = 1000$, $T = 10$).

Discussion

I derived rules-of-thumb for designing reaction-norm studies with sufficient accuracy, precision and power. The results confirm and contrast some recommendations of previous studies, as well as generate some novel suggestions that are of specific interest when designing reaction-norm studies on wild populations. Before discussing these rules-of-thumb, I would like to stress that there are many ways in which the data structure can vary, as exemplified by the different scenarios considered in this study. And for each scenario, I only explored a limited part of the possible parameter space (but reassuringly most results were robust). Moreover, depending on the biological question, one might aim for other levels of accuracy, precision and power than used here. Thus, biologists designing reaction-norm studies are advised to perform their own simulation studies suited to their specific biological context and the R-package *odprism* that I developed should make this job much easier.

SAMPLING MANY INDIVIDUAL OR MANY REPLICATES?

My recommendation to sample at least 40 individuals is higher than Snijders & Bosker (1999), lower than Maas & Hox (2004a), but comparable to what Kreft (1996) advised. In contrast to the recommendation by Martin *et al.* (2011) to choose a ratio of I/J of about a half, my results suggests that it is better to samples more individuals than replicates ($I/J \geq 1$) if one is interested in maximizing the accuracy and precision of variance component estimates. Moreover, random regression models are data hungry, and I suggest that total sampling effort should be at least 1000, comparable to what Kreft (1996) and Martin *et al.* (2011) advised.

A comparison of previous studies that have used random regression approaches confirms that total sample size is a good predictor for statistical evidence for individual variation in slopes (Fig. 6), with the 15 analyses reporting no evidence having a median sample size of only 250, while the 23 analyses reporting positive evidence having a median sample size of 649. Strikingly, the vast majority of analyses with low total sample size concerned behavioural studies, with 23 of 25 behavioural analyses having $N < 1000$, while in only four of 13 analyses on life-history or morphological traits sample size was lower than 1000 (Fig. 6). The risk of false negatives and

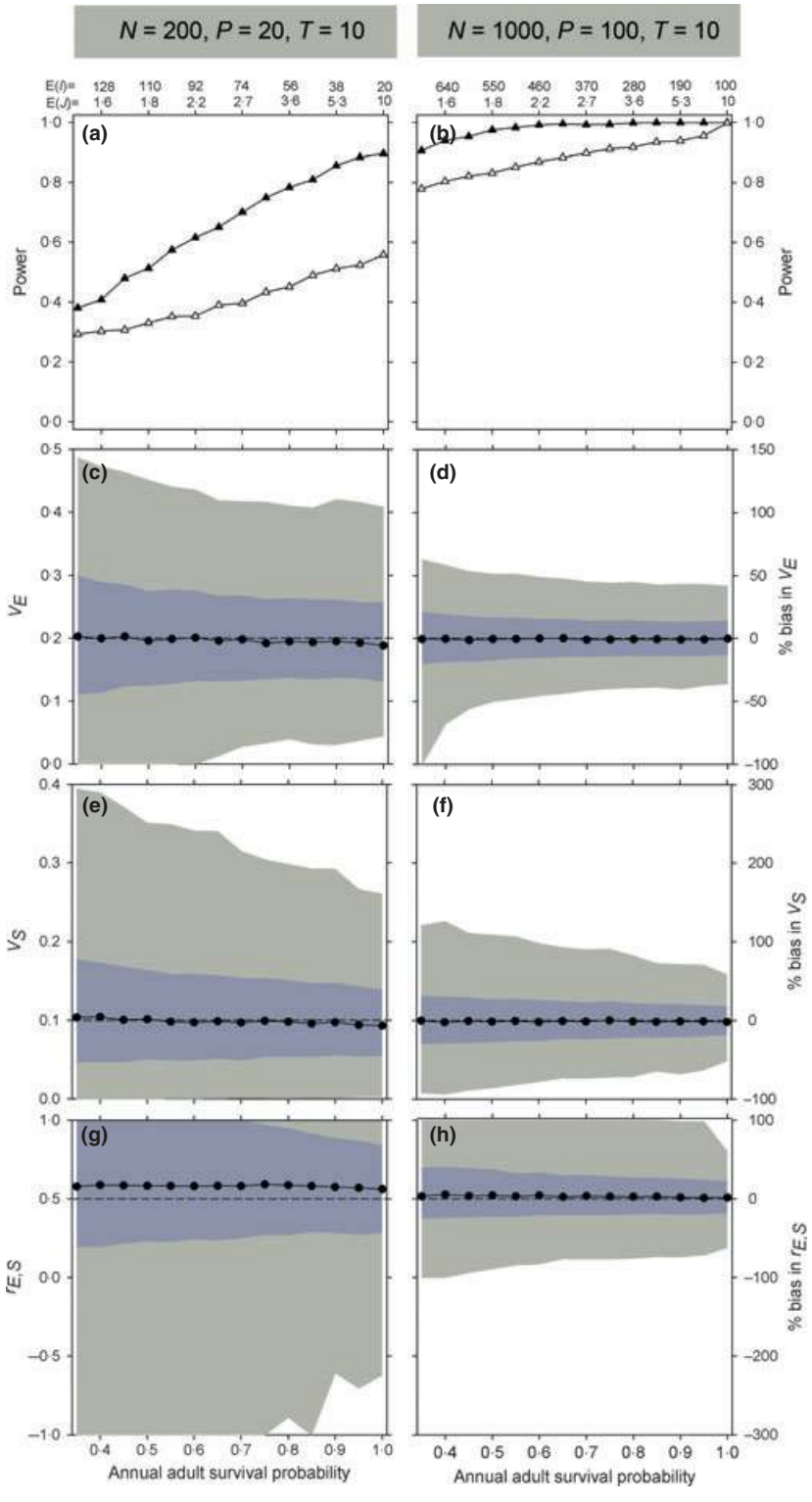

Fig. 5. (a,b) Power of detecting individual heterogeneity in elevation (black triangles) and slopes (white triangles) and accuracy and precision of V_E, (c,d) of V_S (e,f) and of $r_{E,S}$ (g,h) for species with different annual survival probabilities (lower *x*-axes) in scenarios where either a population of 20 individuals (left panels) or 100 individuals (right panels) was followed for a period of 10 years. The expected number of individuals I and replicates per individual J associated with different levels of annual survival are shown on the upper *x*-axes (see eqn 2). In (c–h), circles represent the median estimates of 5000 simulated data sets, while the dark grey and light grey areas depict, respectively, the 25–75% and 2·5–97·5% distribution of the parameter estimates. The horizontal dashed lines depict the values used to generate the data (V_E = 0·2, V_S = 0·1, $r_{E,S}$ = 0·5). Note that the right *y*-axes depict the relative deviation of from the value used to generate the data.

biased or imprecise variance component estimates is thus likely to be greatest in the existing behavioural reaction-norm literature.

The demands on the optimal sampling design in terms of sufficient accuracy, precision as well as statistical power were largely driven by the high sampling effort required for obtaining precise estimates of the correlation between individuals' elevations and slopes. I think that in most biological contexts estimating the correlation parameter precisely is of similar importance as estimating the variances in elevations and slopes precisely, because the correlation parameter directly affects the patterns of variation in individual reaction norms (i.e. the amount of fanning in or out; see fig. 2 in Nussey, Wilson & Brommer 2007).

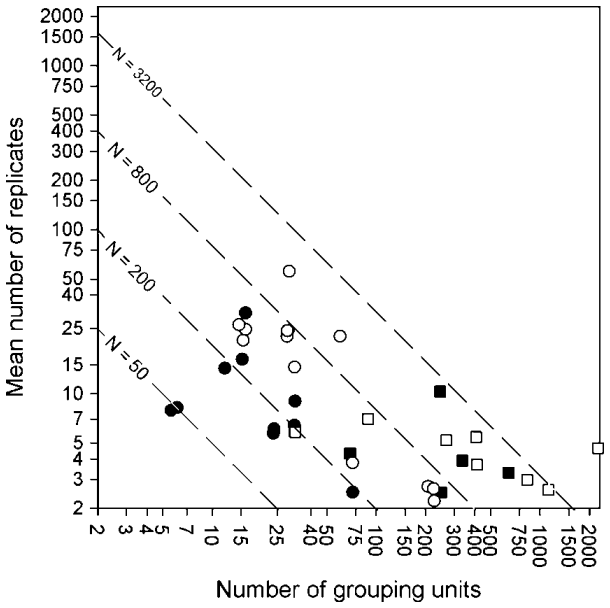

Fig. 6. Sampling design of 38 published analyses (see Appendix S1 for details) that used random regression models to test for variation in slopes of behavioural traits (circles) or of life-history/morphological traits (square symbols). Analyses that reported evidence for variation in slopes among grouping units are depicted by white symbols, while analyses that reported no evidence for variation in slopes are in depicted in black.

TYPE OF ENVIRONMENTAL VARIABILITY

Environmental variables that are either the same or different for all individuals at a given sampling occasion are both common in the reaction-norm literature (Appendix S1). I showed that the type of environmental variability does not strongly affect sampling design, except that when all individuals experience the same environmental conditions, one should strictly avoid sampling all individuals only twice (Fig. 4), even though this is the simplest possible design for a reaction-norm study. These results sharply contrast to the situation when all individuals experience different environmental conditions at a given sampling occasion, in which case sampling all individuals twice does not typically result in identifiability problems (Fig. 3). Notwithstanding, I have noticed that biologists–possibly because of their familiarity with classical regression models – seem to be averse towards applying random regression models to individuals that have been measured only twice. Such hesitation is unjustified as long as one samples individuals under variable conditions. However, sampling individuals more than twice is of course crucial for assessing the nonlinearity of reaction norms.

Many studies with unbalanced data sets have removed all individuals that have been measured only once, twice or even three times prior to analyses (e.g. Brommer *et al.* 2005; Nussey *et al.* 2005; Reed *et al.* 2006; Charmantier *et al.* 2008; Dingemanse *et al.* 2011). Clearly, removing individuals measured only twice or thrice is a waste of valuable information. Even including individuals measured only once in the analyses can be valuable, as they can improve the estimation of non-slope model parameters and thereby indirectly also increase the statistical power to detect individual variation in slopes (Martin *et al.* 2011).

CONSTRAINTS IMPOSED BY THE STUDY SPECIES

I showed that sampling effort should be substantially higher for short-lived than for long-lived species when investigating annual traits, all else being equal (Fig. 5, Figs F1 and F2 in Appendix S6). The main problem of sampling annual traits in short-lived species is that most individuals can only be sampled once and then die, and that such individuals do not directly contribute to estimation of individual variation in plasticity. However, studying short-lived species may also have its benefits in reaction-norm studies: often the reason for quantifying random slopes is that evidence for 'I × E' provides the basis for further investigations into genotype by environment interactions ('GxE'; Nussey, Wilson & Brommer 2007). To quantify 'GxE', one typically needs deep pedigrees, which take much less years to collect for species with a short generation time. In fact, some of the best-known examples of reaction-norm studies in the wild are on very short-lived species (Brommer *et al.* 2005; Nussey *et al.* 2005; Charmantier *et al.* 2008); notably, each of these studies had considerable sampling effort ($T \geq 23$ years, $N > 2000$).

TRAITS UNDER VIABILITY SELECTION

Despite the fact that survivors can be a nonrandom subset of the population with respect to plastic traits (Vaupel, Manton & Stallard 1979; van de Pol & Verhulst 2006; Hadfield 2008), reassuringly, variance components were barely affected by directional viability selection on traits in the simulations. Apparently, the combination of viability selection against individuals with specific trait values and their replacement by individuals with random trait values barely affect the among-individual variance components. Possibly, the situation is different when viability selection takes other forms (e.g. divergent), or acts not on the trait value, but instead acts directly on the slopes of individuals, as might occur in cases of phenological mismatch (Nussey *et al.* 2005). Also, viability selection might more strongly impact variance components when there is a heritable component to an individual's elevation or slope.

SIMULATIONS AS A BEST CASE SCENARIO

In several respects, the situation in the real world is likely to be more complex and thus, simulations represent a best case scenario. The rules-of-thumbs derived here should therefore be interpreted conservatively. For example, in reality, there are unobserved variables that cause additional unexplained heterogeneity. Also, reaction norms might be nonlinear, which potentially alters the optimal design considerably. Model assumptions concerning normality or homogeneity of variance might be violated by the data (Maas & Hox 2004a,b). And because binary and count data contain less information than normal data, studies investigating noncontinuous traits almost

certainly require higher sampling effort for accurate and precise estimation (Moineddin, Matheson & Glazier 2007). Finally, my study did not consider the case of nonindependence among individuals due to genetic relatedness; results may thus not be representative for random regression animal models, because such models not only estimate additional variance components but also incorporate more information (i.e. pedigree).

CAUTION IN INTERPRETING RANDOM SLOPES AS I × E

In the reaction-norm literature, evidence for individual variation in slopes is typically interpreted as evidence for the existence of an individual by environment interaction ('I × E'; Nussey, Wilson & Brommer 2007). However, evidence for individual variation in slopes may also result from characteristics that are not necessarily 'consistently' different between individuals. For example, suppose that the response to the environment changes with age ('A × E'; van de Pol, Osmond & Cockburn, in press) and one collects data on a population consisting of individuals of variable age. In this scenario, the fact that one measured some individuals at older ages than others may result in the slope of some individuals to be different from others. However, such individual variation in slopes is not caused by I × E but instead caused by unobserved A × E. Thus, state variables may interfere with interpreting variance in slopes as evidence for I × E, and it therefore is important to account for key state variables (e.g. age, body size, social status) in reaction-norm studies.

Acknowledgements

The Australian Research Council supported my research with a postdoctoral fellowship (DP1092565). I would like to thank Nicolas Margraf and Lyanne Brouwer for discussion and Julien Martin, Nigel Yoccoz and an anonymous reviewer for helpful comments.

References

Bates, D.M., Maechler, M. & Dai, B. (2008) lme4: linear mixed-effects models using S4 classes. URL http://lme4.r-forge.r-project.org/ [version 0.999375-40]

Brommer, J.E., Merilä, J., Sheldon, B.C. & Gustafsson, L. (2005) Natural selection and genetic variation for reproductive reaction norms in a wild bird population. *Evolution*, **59**, 1362–1371.

Charmantier, A., McCleery, R.H., Cole, L.R., Perrins, C.M., Kruuk, L.E.B. & Sheldon, B.C. (2008) Adaptive phenotypic plasticity in response to climate change in a wild bird population. *Science*, **320**, 800–803.

Chevin, L.-M., Lande, R. & Mace, G.M. (2010) Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. *PLoS Biology*, **8**, e1000357.

Dingemanse, N.J., Kazem, A.J.N., Réale, D. & Wright, J. (2010) Behavioural reaction norms: animal personality meets individual plasticity. *Trends in Ecology & Evolution*, **25**, 81–89.

Dingemanse, N.J., Bouwman, K.M., van de Pol, M., van Overveld, T., Patrick, S., Matthysen, E. & Quinn, J.S. (2011) Variation in personality and behavioural plasticity across four populations of the great tit *Parus major*. *Journal of Animal Ecology*, doi: 10.1111/j.1365-2656.2011.01877.x

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

Henderson, C.R. (1982) Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions. *Biometrics*, **38**, 623–640.

Hox, J.J. (2010) *Multilevel Analysis. Techniques and Applications*. Routledge, New York.

de Jong, G. (1990) Quantitative genetics of reaction norms. *Journal of Evolutionary Biology*, **3**, 447–468.

Kreft, I.G.G. (1996) *Are Multilevel Techniques Necessary? An Overview Including Simulation Studies*. California State University, Los Angeles.

Maas, C.J.M. & Hox, J.J. (2004a) Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, **58**, 127–137.

Maas, C.J.M. & Hox, J.J. (2004b) The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, **46**, 427–440.

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

Moineddin, R., Matheson, F.I. & Glazier, R.H. (2007) A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, **7**, 34.

Nussey, D.H., Wilson, A.J. & Brommer, J.E. (2007) The evolutionary ecology of individual phenotypic plasticity in wild populations. *Journal of Evolutionary Biology*, **20**, 831–844.

Nussey, D.E., Postma, E., Gienapp, P. & Visser, M.E. (2005) Selection on heritable phenotypic plasticity in a wild bird population. *Science*, **310**, 304–306.

Pigliucci, M. (2001) *Phenotypic Plasticity: Beyond Nature and Nurture*. Johns Hopkins University Press, Baltimore.

Pinheiro, J.C. & Bates, D.M. (2000) *Mixed-Effects Models in S and S-plus*. Springer Verlag, New York.

van de Pol, M. (2011) R-package *odprism:* Optimal Design and Performance of Random Intercept and Slope Models. URL http://cran.r-project.org/web/packages/odprism/ [version 1.0]

van de Pol, M. & Verhulst, S. (2006) Age-dependent traits: a new statistical model to separate within- and between-individual effects. *American Naturalist*, **167**, 766–773.

van de Pol, M. & Wright, J. (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. *Animal Behaviour*, **77**, 753–758.

van de Pol, M., Osmond, H. & Cockburn, A. (In press) Fluctuations in population composition dampen the impact of phenotypic plasticity on trait dynamics in superb fairy-wrens. *Journal of Animal Ecology*. doi: 10.1111/j.1365-2656.2011.01919.x.

R Development Core Team (2010) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna. http://www.r-project.org [version 2.11.0].

Raudenbush, S.W., Spybrook, J., Congdon, R., Liu, X. & Martinez, A. (2011) Optimal Design software for multi-level and longitudinal research. URL www.wtgrantfoundation.org or from sitemaker.umich.edu/group-based.

Reed, T.E., Wanless, S., Harris, M.P., Frederiksen, M., Kruuk, L.E.B. & Cunningham, E.J.A. (2006) Responding to environmental change: plastic responses vary little in a synchronous breeder. *Proceedings of the Royal Society B: Biological Sciences*, **273**, 2713–2719.

Reed, T.E., Waples, R.S., Schindler, D.E., Hard, J.J. & Kinnison, M.T. (2010) Phenotypic plasticity and population viability: the importance of environmental predictability. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 3391–3400.

Scheipl, F., Greven, S. & Kuchenhoff, H. (2008) Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, **52**, 3283–3299.

Scherbaum, C.A. & Ferreter, J.M. (2009) Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, **12**, 347–367.

Schielzeth, H. & Forstmeier, W. (2009) Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology*, **20**, 416–420.

Sih, A., Bell, A.M., Johnson, J.C. & Ziemba, R.E. (2004) Behavioral syndromes: an integrative overview. *The Quarterly Review of Biology*, **79**, 241–277.

Snijders, T.A.B. (2005) Power and sample size in multilevel linear models. *Encyclopedia of Statistics in Behavioral Science, volume 3* (eds B.S. Everitt & D.C. Howell), pp. 1570–1573. Wiley, Chicester.

Snijders, T.A.B. & Bosker, R.J. (1993) Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, **18**, 237–259.

Snijders, T.A.B. & Bosker, R.J. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE Publications, London.

Vaupel, J.W., Manton, K.G. & Stallard, E. (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439–454.

Verbeek, M. (2000) *A Guide to Modern Econometrics*. Wiley, New York.
Vindenes, Y., Engen, S. & Saether, B.-E. (2008) Individual heterogeneity in vital parameters and demographic stochasticity. *The American Naturalist*, **171**, 455–467.
Visscher, P.M. (2006) A note on the asymptotic distribution of likelihood ratio tests to test variance components. *Twin Research and Human Genetics*, **9**, 490–495.
Visser, M.E. (2008) Keeping up with a warming world; assessing the rate of adaptation to climate change. *Proceedings of the Royal Society B: Biological Sciences*, **275**, 649–659.

Received 21 June 2011; accepted 21 September 2011
Handling Editor: Nigel Yoccoz

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Overview of studies that used random regression models to estimate individual variation in slopes.

Appendix S2. Accuracy, precision and power of random regression models as a function of the number of individuals and replicates sampled for various parameter conditions, in a situation where all individuals experience different environmental conditions at a given sampling occasion.

Appendix S3. Accuracy, precision and power of random regression models as a function of the number of individuals and replicates for various parameter conditions, in a situation where all individuals experience the same environmental conditions at a given sampling occasion.

Appendix S4. Accuracy, precision and power of random regression models as a function of the number of individuals for different types of environmental variability.

Appendix S5. Accuracy, precision and power of random regression models as a function of annual survival (i.e. species longevity) for various parameter conditions.

Appendix S6. Accuracy, precision and power of random regression models as a function of species longevity and the size P and the number of years T a population is sampled.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.