Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 14–24 doi: 10.1111/j.2041-210x.2012.00251.x

Nested by design: model fitting and interpretation in a mixed model era

Holger Schielzeth[1]* and Shinichi Nakagawa[2]

[1]*Department of Evolutionary Biology, University of Bielefeld, Morgenbreede 45, 33615, Bielefeld, Germany; and* [2]*National Centre of Growth and Development, Department of Zoology, University of Otago, 340 Great King Street, Dunedin 9054, New Zealand*

Summary

1. Nested data structures are ubiquitous in the study of ecology and evolution, and such structures need to be modelled appropriately. Mixed-effects models offer a powerful framework to do so. Nested effects can usually be fitted using the syntax for crossed effects in mixed models, provided that the coding reflects implicit nesting. But the experimental design (either nested or crossed) affects the interpretation of the results.

2. The key difference between nested and crossed effects in mixed models is the estimation and interpretation of the interaction variance. With nested data structures, the interaction variance is pooled with the main effect variance of the nested factor. Crossed designs are required to separate the two components. This difference between nested and crossed data is determined by the experimental design (thus by the nature of data sets) and not by the coding of the statistical model.

3. Data can be nested by design in the sense that it would have been technically feasible and biologically relevant to collect the data in a crossed design. In such cases, the pooling of the variances needs to be clearly acknowledged. In other situations, it might be impractical or even irrelevant to apply a crossed design. We call such situations naturally nested, a case in which the pooling of the interaction variance will be less of an issue.

4. The interpretation of results should reflect the fact that the interaction variance inflates the main effect variance when dealing with nested data structures. Whether or not this distinction is critical depends on the research question and the system under study.

5. We present mixed models as a particularly useful tool for analysing nested designs, and we highlight the value of the estimated random variance as a quantity of biological interest. Important insights can be gained if random-effect variances are appropriately interpreted. We hope that our paper facilitates the transition from classical ANOVAs to mixed models in dealing with categorical data.

Key-words: ANOVA, categorical data, experimental design, hierarchical models, interaction variance, mixed-effects models, variance components analysis

Introduction

Answering ecological and evolutionary problems often requires complex data with multiple predictors for a response variable of interest. Multiple categorical variables can be either nested or crossed depending on the experimental design employed during data collection (Scheiner & Gurevitch 2001; Quinn & Keough 2002; Ryan 2007; Hinkelmann & Kempthorne 2008; Kirk 2009). Furthermore, data might be structured in a variety of ways, which often requires appropriate control for random effects (Bolker *et al.* 2009; Zuur, Ieno & Elphick 2010). We here discuss nested and crossed data structures with the aim of assisting the biological interpretation of statistical models. We also highlight the value of mixed-effect models as a powerful tool for modelling nested data sets.

This paper is primarily concerned with study designs that encompass two categorical predictors (called factors or factorial predictors in the following), although our points also apply to study designs with more than two categorical variables. In a nested design, each level of the nested predictor is uniquely associated with only one level of the higher-level predictor (Table 1, Fig. 1). In a crossed (or factorial) design, at least one level of each predictor is associated with more than one level of the other predictors (Table 1, Fig. 1). Crossed designs can further be separated into partially or fully crossed. In a fully crossed (or full-factorial) design, there are observations for all combinations of levels of the two predictors, that is, each level of one predictor is associated with each level of the other predictors, while in a partially crossed design, some combinations of the two predictors have not been sampled (Table 1). This basic categorization of the study design does not depend on whether factors are fitted as fixed or random effects, a distinction that we will discuss in more detail in the second half of the paper.

Data sets with nested data structures are very common in evolutionary and ecological studies (Quinn & Keough 2002;

*Correspondence author. E-mail: holger.schielzeth@uni-bielefeld.de

Table 1. Schematic illustration of crossed and nested designs

Nested design				
	Factor 2			
Factor 1	a	b	c	d
A	X	X		
B			X	X
Partially crossed design				
	Factor 2			
Factor 1	a	b	c	d
A	X	X		
B	X	X		
C			X	X
D			X	X
Fully crossed design				
	Factor 2			
Factor 1	a	b	c	d
A	X	X	X	X
B	X	X	X	X
C	X	X	X	X
D	X	X	X	X

Factor combinations for which observations are available are marked with crosses. Factor 1 has 2–4 levels indicated by upper-case letters, while Factor 2 has four levels indicated by lower-case letters. Two factors can be nested (top), partially crossed (middle) or fully crossed (bottom), depending on whether all levels of the nested factor (here 'Factor 2') are uniquely associated with the higher-level factor ('Factor 1', nested designs) or if all (fully crossed design) or at least some (partially crossed design) more combinations of factor levels were sampled.

Kéry 2010). Our main aim is to show that the key consequence of nesting is the pooling of the interaction variance with the main effect variance of the nested factor. Nested designs are more constrained than crossed (factorial) designs in that the former does not allow the separate estimation of interaction variance (Ryan 2007), while a crossed design does allow such estimation. Nested designs can be fitted in a classical linear model with nested fixed effects, but this task is not entirely trivial, because the degrees of freedom need to be adapted to the experimental design in order to get the correct mean sum of squares (Underwood 1997; Quinn & Keough 2002; Gelman 2005). Such data can often be more easily analysed in a mixed-(effects) model that estimates the standard errors appropriately (Gelman 2005).

Our secondary aim in this paper is therefore a discussion of mixed models as a powerful tool for modelling structured data. Mixed models are also known as hierarchical or multilevel models in the social and medical sciences (Congdon 2007; Gelman & Hill 2007; Goldstein 2011; Snijders & Bosker 2011). Mixed models feature random effects that allow clustering of data in groups. The distinguishing characteristic of random effects is the explicit modelling of the between-group variance using a hyperparameter(s) (sensu Gelman & Hill 2007; see below and Table 2). Fixed and random factors can be nested or crossed with each other, depending on whether some factor varies only within levels of another factor (i.e. nested) or whether the levels at which two factors vary are independent of each other (i.e. crossed). With mixed-effects models, two categorical predictors can be fitted using the syntax for crossed effects even if the design is nested. Implicit nesting through appropriate coding (as discussed in section Nesting and study design) ensures that the design matrices are built correctly and that the uncertainty of the fixed effects is estimated appropriately. Again, the key difference between nested and crossed designs lies in the interpretation of the variance components that is inflated by the interaction variance for the nested factor in a nested design.

Nesting and study design

We distinguish two types of nesting. Sometimes factors are naturally nested. For example, dry biomass of a primary producer might have been measured at different sites. Biomass was quantified in each of multiple plots within each study site, and two extraction replicates were taken per plot. Plots are nested in study sites and extractions constitute replicates within plots. It would be impossible to break the nesting of plots within study sites, because it is infeasible to translocate plots to different sites. It would even be irrelevant to break the nesting of plots within study sites, because each plot could not have existed at different sites. The two spatial scales are biologically nested, and if we aim to understand the biological system, it is irrelevant to imagine crossing sites and plots.

In other cases, the effects are nested by study design. For example, we might be interested in whether supplementary feeding of adult birds influences the size of the offspring. Supplementary feeding (the treatment) might have been applied to egg-laying females, and after hatching, multiple chicks might have been measured within individual broods. Females are nested within treatments, and chicks are nested within females (and clustered within broods). It might have been possible to design the study with some of the factors crossed. For example, different treatments could be applied to the same females in different laying cycles, so that female identities are crossed with treatment. Furthermore, eggs could have been transferred among clutches so that after hatching, each brood contains chicks from multiple females that have experienced different treatments. Cross-fostering would break the nesting of chicks within broods, which is important if one wants to distinguish between an effect of the treatment before/during egg laying (maternal effects) and an effect on post-laying incubation and parental care (Mousseau & Fox 1998). A crossed design would have been feasible and is biologically relevant. Therefore, the data are nested by design and not naturally nested. The differences between natural nesting and nesting by design will become important when it comes to the interpretation of the variance explained by the nested factor.

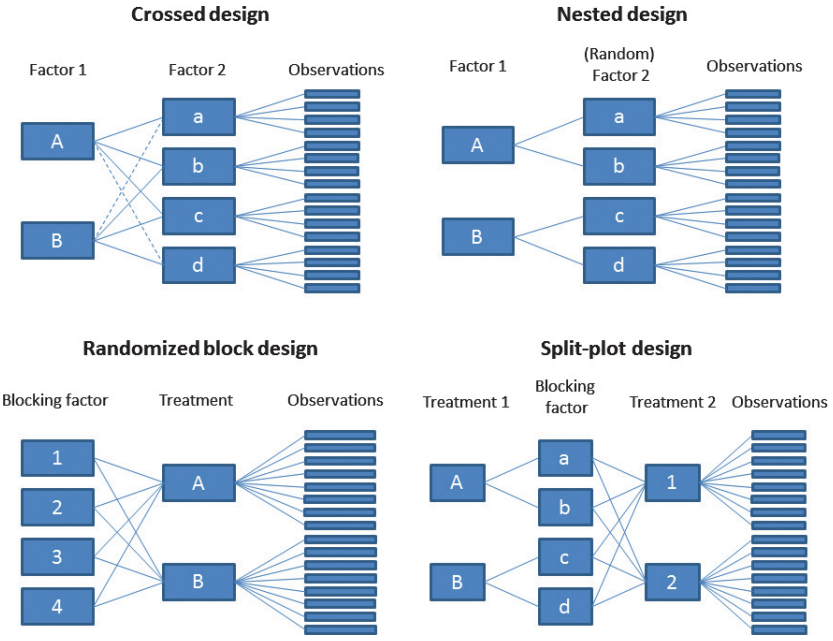

Fig. 1. Schematic illustrations of four classical study designs. A partially (solid lines only) or fully (solid and dashed lines) crossed design allows the estimation of main effects for the two factors and of the interaction variance. In a nested design, the nested factor is typically conceptually random, even though it might be fitted as a fixed effect (Factor 1 is a group-level predictor relative to Factor 2). In a randomized block design, the blocking factor is typically conceptually random, even though it might be fitted as a fixed effect (Treatment 2 is a data-level predictor). The block × treatment interaction variance can be estimated if there are replicate observations for each block–treatment combination. A split-plot design combines group-level and data-level predictors (see Fig. 3). The block × treatment 2 and the treatment 1 × treatment 2 interaction variances, but not the block × treatment 1 interaction variance, can be estimated. Factor levels are labelled by upper-case, lower-case letters or numbers.

There are two ways of coding the levels of a nested factor (that we label 'Factor 2' in the following). One way of coding is to uniquely label levels within levels of Factor 1 (the factor in which Factor 2 is nested). This might be done for convenience in the data collection. For example, it is easy to keep track of the 10 plots within each of multiple study sites, which would result in plots being labelled 1–10 in each of the study sites, such that different plots at different study sites take the same label even though they are geographically distinct. While this is a common habit, it is prone to misinterpretations and we strongly advise against it. Instead, levels of the nested factor should be labelled uniquely within the whole data set. For example, in the nested design of biomass study described above, we should label all plots and each extracting replicate uniquely. After all, each plot is a unique entity. If we have investigated 10 plots in 10 study sites, their identifiers should run 1–100 (or otherwise unique) rather than recycling identifiers 1 and 10 for each site. A main advantage of this way of coding is that it implicitly describes a nested data structure that a computer programme or a colleague would recognize the structure as nested without further explanations. If plots are labelled uniquely, nobody would ask whether plot 1 and plot 11 are the same, but if plots were merely labelled uniquely within study sites, it needs additional information that is not coded in the data set, clarifying that plot 1 at study site 1 is not the same as plot 1 at study site 2.

Nested and crossed: the key difference

In a two-way factorial design (such as a classical two-way ANOVA scenario), there are four biological sources of variance that can potentially be estimated (Fig. 2). For example, a supplementary feeding treatment (two levels) might have been applied to pairs of a bird species at 10 different study sites (10 levels) with both feeding treatments applied to 10 pairs each at each of the 10 sites. The response variable of interest is the number of fledglings produced. This essentially is a 2 × 10 full-factorial design with balanced sampling ($N = 200$). The four variance components that can be estimated are as follows:

1 Main effect (marginal) variance explained by Factor 1: This is the variance in the response explained by the supplementary feeding treatment averaged across the 10 study sites.

2 Main effect (marginal) variance explained by Factor 2: This is the variance in the response explained by the 10 study sites averaged across the two supplementary feeding treatments.

3 Interaction variance explained by factor combinations: This is the variance in the response explained by the specific combinations of the treatment × site after controlling for the average effect of the supplementary feeding treatment across all sites and the average effect of the study sites across the two treatments.

4 Residual variance: This is the variance in the response that is unexplained by treatment, study site and their interaction and hence the variance that remains within cells, that is, the variance among pairs in the number of fledglings after accounting

Table 2. Operational definitions of key terms

Term	Explanation
Data (unit) level	The level of individual observations (data, units of the analysis) and the most basic or lowest level. The unexplained variability at the data level is expressed as the residual variance
Data-level predictor	Explanatory or independent variable that varies at the data level, such that different observations take different values independent of any grouping level
Factor	Categorical predictor (that can be fitted as a fixed or random effect)
Fixed effect	Effects that are estimated at each factor level independently of all other factor levels, that is, only observations within each level contribute to the estimate. Factors can be fitted as fixed effects, but can still be conceptually random in the sense that they represent a random sample of levels rather than distinct treatments (e.g. block effects)
Group-level predictor	Explanatory variable that varies at the grouping level, such that all observations within the same group take the same value
Grouping level	Clusters of observations which constitute a hierarchical level above the data level. For example, individuals (data level: replicate observations per individual) or groups of individuals (data level: single observations per individual)
Groups	Used in the statistical sense of any grouping (or clustering) of related observations. For example, individuals, species, blocks, plots
Hyperparameter	An estimator at a higher hierarchical level that controls estimates at the group level. In classical mixed models, the group-level variance is a hyperparameter that estimates the variance of group-level means, which are themselves parameters of the model. Both the group-level variance and the group-level means are estimated from the data. (The term 'hyperparameter' has a second meaning in a Bayesian context that differs from the definition given here.)
Main effect/marginal effect	The effect of a categorical predictor on the response when moving from one treatment level to another while holding all other predictors constant. The constancy of the marginal effect across values of other predictors distinguishes marginal effects from interaction effects that vary conditional on values of (one or more) other predictors
Random effect	Effects that are estimated at each factor level, but where the distribution of the estimates is explicitly modelled by hyperparameters. The variance of the random effects can be considered the 'unexplained' variance at this level in the sense that the detailed causes of such random-effect variance are unknown. Estimates are influenced by shrinkage towards the population mean
Random slopes	Most random effects that are fitted in mixed models are random intercept effects, that is, mean response value are allowed to vary among groups. Random slopes represent an interaction between a fixed factor and a random factor. Significant random-slope variance means that the magnitude of the between-group variance varies with values of a covariate or, equivalently, that the effect of a covariate varies among groups
Shrinkage	A property of random-effect estimation in mixed models. Group means are not only influenced by observations from a particular group, but also by the population mean, such that the random-effect estimates for each group are closer to the population mean than the mean of the observations from a particular group (i.e. they are 'shrunken' towards the population mean). The effect is more pronounced for groups with a small number of observations
Treatment	An experimental manipulation that is of primary interest in a study. We use the term in a wider sense, including also factors that are not under direct control of the experimenter (e.g. breeding status of an individual), therefore covering also quasi-experimental designs (Ryan 2007). Treatments will typically be fitted as fixed factors

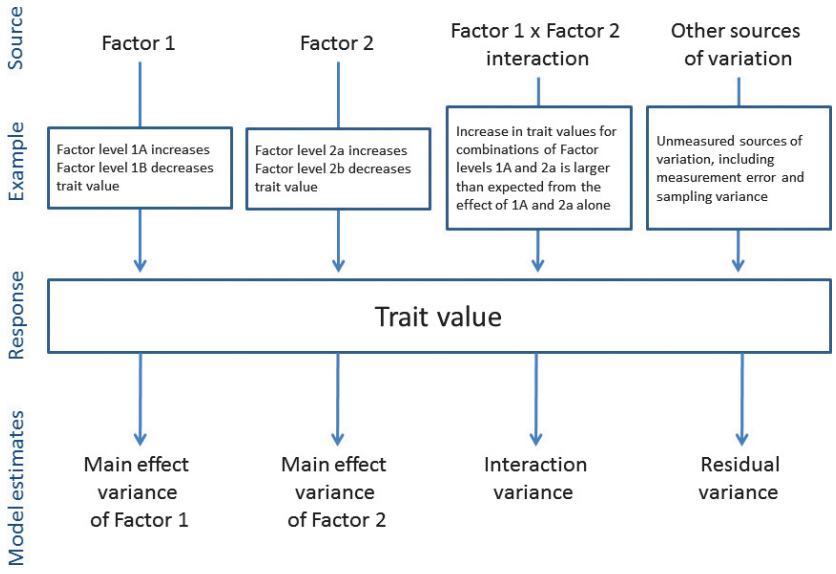

Fig. 2. Biological sources of variance that affect the trait value of interest and how they can be estimated in a crossed experimental design. Nested sampling cannot change the biological sources of variance (they are part of the biological reality), but the sampling design affects the way in which sources of variance can be estimated from the data (see text for details). Levels of Factor 1 are labelled by upper-case letters, and levels of Factor 2 are labelled by lower-case letters.

for the mean number of fledglings in each combination of factor levels.

In the crossed design that we describe above, it is possible to estimate all four variance components (Table 3). If we decide to remove the interaction term from the model, because we consider it of little biological interest, then the interaction variance will not be estimated by the model. However, biological sources of interaction variance cannot disappear because the total variance in the response values is unaffected by our modification to the model. As a consequence, the removal of the interaction term will lead to an increase in the residual variance component, because the interaction variance will be pooled with the residual variance (Table 3). The two sources of variation are thus conflated.

If we had applied a nested sampling design, such application design would affect how the variances can be estimated. For example, we might have applied the treatment to complete study sites so that 20 pairs in each of five sites received supplementary feeding, while 20 pairs in each of the five other sites received a control treatment. We might be constrained to such a treatment, if the species is colonial so that we are not able to apply the treatment to individual pairs, but only to the colony as a whole. However, biological sources of variation still remain unchanged. There would still be an average effect of the treatment across all sites and variation in the number of fledglings between sites. The interaction variance between site and treatment will also be present in the biological system if it matters which treatment is applied to which site, for example because there is natural variation in food availability between sites and a supplementary feeding in a rich environment has a smaller effect than a supplementary feeding in a poor environment. In this example, we had to choose which treatment we would apply to which site, and thus, we do not know what would have happened if we had applied the other treatment. So where is the interaction variance? Indeed, the interaction effect and the nested factor main effect variance are completely confounded. Therefore, the interaction variance will be pooled (conflated) with the main effect variance of the study site variance. We cannot decide if the between-site variation that we estimate is caused by average differences among sites (e.g. productivity differences) and/or by differences among sites in the effect of the treatment. The estimate that we obtain for the between-site variance is inflated by the interaction variance.

This is a fundamental difference between a nested design and a crossed design without explicit modelling of the interaction term. In both cases, we will get estimates of three variance components, but in one situation (crossed design without interaction), the interaction variance will increase the residual variance, while in the other (nested design) the interaction variance will inflate the main effect variance of the nested factor (Table 3). The key difference between the two designs therefore lies in the estimation of the interaction variance (Table 3).

This point relates to a well-known issue in model fitting, namely the interpretation of interaction effects (Aiken & West 1991; Engqvist 2005). When data from a crossed design are analysed in a model with an interaction term, but with one of the main effects removed, then the interaction term becomes very difficult to interpret. The estimate for the interaction term no longer represents the interaction variance only but is conflated with the main effect variance of the removed factor (Table 3). For example, if we have a full-factorial design that includes age class, treatment and their interaction, but we were to remove age from the model while keeping the age × treatment interaction term, the estimate for the interaction will then be inflated by the main effect variance of age. The pooling of the main effect variance and the interaction variance is the reason for the well-appreciated warning against the removal of main effects in the presence of interactions (Aiken & West 1991).

A word of caution is required for the cases of crossed but unbalanced designs. In such cases, the separation of the variance will be less precise, and the more so, the more unbalanced the design. Balanced sampling increases the power to separate variance components and is therefore an important aim in data

Table 3. Sources of variance and how they are estimated

Model set-up (fixed and random effects)			Study design		
Source of variation	One fixed and one random	Two random	Crossed	Crossed (without interaction)	Nested
Main effect of Factor 1	Variance of fixed factor	Variance of random Factor 1	V_1	V_1	V_1
Main effect of Factor 2	Variance of random factor (random-intercept variance)	Variance of random Factor 2	V_2	V_2	$V_{12} + V_2$
Interaction	Interaction variance between fixed and random factor (random-slope variance)	Interaction variance between random Factors 1 and 2	V_{12}	–	–
Residual	Residual variance	Residual variance	V_R	$V_{12} + V_R$	V_R

In a two-way factorial design, there are the four sources of variance: the main effect variance of Factor 1 (V_1), the main effect variance of Factor 2 (V_2), the interaction variance (V_{12}) and the residual variance (V_R). Whether or not the four variances can be separated depends on the study design and how they are estimated depends on the set-up of the model.

collection. But unbalanced designs are common in ecology and evolution, in particular in field studies where it is often difficult to ensure perfect balance. Therefore, mixed models, which provide appropriate estimators, are particularly valuable for such unbalanced data sets.

Classical experimental designs

We here discuss a selection of classical experimental designs that readers might be familiar with. The section aims to show that classical experimental designs can be conceptualized in a more general linear modelling framework, rather than being something distinctly different.

Classical examples of nested and crossed experimental designs are the breeding designs employed for quantitative genetic analyses (Comstock & Robinson 1952). The North Carolina I design is a typical example of a nested design. Sires are mated to multiple dams that produce multiple offspring. Observations are clustered in dams, and dams are nested within sires. To put it another way, full-sib families are nested within half-sib families. Such data are nested by design, because an alternative breeding design, the North Carolina II design, could have been applied. In a North Carolina II design, sires are mated to multiple dams and dams are mated to multiple of these sires. The North Carolina II design can be fully or partially crossed (Lynch & Walsh 1998). Unlike the nested North Carolina I design, the crossed North Carolina II design allows the separate estimation of maternal effect and dominance variance that are conflated with the dam variance in the North Carolina I design (Falconer & Mackay 1996; Lynch & Walsh 1998).

Nested and (partially or fully) crossed effects can be considered the basic units of more complex models. In a randomized block design, treatments are applied to randomly selected blocks (Fig. 1). The design is either partially or fully crossed, depending on whether all or only some of the treatment levels are applied to each block (randomized complete block designs or generalized block designs, Quinn & Keough 2002; Kirk 2009; Rasch *et al*. 2011). A block design is therefore a classical crossed design with one of the factors being a random factor and treatment being a data-level predictor (see discussion below). If there are replications within block–treatment combinations, it is possible to estimate the block × treatment interaction variance, that is, whether blocks differ in their treatment effects. Without replications, the interaction variance is pooled with the residual variance; hence, it cannot be estimated.

Split-plot designs combine crossed and nested factors (Quinn & Keough 2002; Fig. 1). There are at least three factors involved, one of them being random and the other two fixed. One treatment is applied to plots (a group-level predictors, see discussion below), while the other is applied in a partially or fully crossed manner to plots (a data-level predictor, see discussion below). The plot × treatment interaction variance is necessarily pooled with the main effect of the group-level predictors, whereas for the data-level predictors, it is possible to estimate the marginal effect and the interaction variance separately, if there is replication of plot × data-level predictor combinations. With only one observation per plot × data-level predictor combination, the plot × data-level predictor interaction variance is pooled with the residual variance.

Another well-known class of experimental designs are repeated-measures designs that are characterized by measurements of each subject on more than one occasion. For example, each subject sequentially experiences at least two conditions or subjects are monitored longitudinally (i.e. longitudinal studies, Singer & Willett 2003). Repeated-measures designs are also referred to as 'crossover' trials (Diaz-Uriarte 2002), and the treatment (a combination of conditions) and subject are fully crossed with each other (Table 1). Traditionally, data from such designs are analysed by repeated-measures ANOVAS, which unfortunately does not allow any missing values. Mixed models with random slopes (i.e. random-slope models) can also be applied to repeated-measures designs with missing data (partially crossed) or without missing data (fully crossed); usually, subjects are modelled as a random (intercept) effect, the treatment as a fixed effect and the temporal/sequential effect of the treatment as random slopes (see Diaz-Uriarte 2002; Schielzeth & Forstmeier 2009).

Mixed-effects models

Mixed-(effects) models are a statistical framework that features fixed and random factors. Mixed models explicitly model hierarchical data structures by clustering observations into groups (Gelman & Hill 2007; Bolker *et al*. 2009). Clustering might be considered a case of nesting, because observations uniquely belong to particular groups (Zuur *et al*. 2009), but we prefer the terms 'structured data', 'grouped data' or 'clustered data' over 'nested data' in this case, because this terminology avoids the confusion with nested factors, and it importantly allows for crossed random effects (Gelman & Hill 2007). Grouping structures might arise from repeated measurements on the same individuals, but also from spatial or temporal structure, family structures, social groups of organisms, etc. At the lowest hierarchical level, there are individual observations. We call this level the data or unit level (Gelman & Hill 2007; Table 2). Individual observations are grouped by random factors. Random factors therefore constitute the grouping level. Because of the modelling of different levels of grouping, mixed models are often called hierarchical or multilevel models, particularly in the social sciences (Goldstein 2011; Snijders & Bosker 2011).

Random factors are predictors where the distribution of individual coefficients is explicitly modelled by hyperparameters (see Table 2), in the typical case by estimating the between-group variance (Gelman & Hill 2007). Unlike fixed factors that are estimated purely based on observations made for a particular factor level, the estimates of random factors are influenced by the population mean; in fact, their estimates are drawn towards the population mean ('shrinkage', see Table 2; McCulloch & Neuhaus 2005; Snijders & Bosker 2011). There exists an extensive discussion about fixed and random effects (see below). In practical applications, variables are modelled as random effects if the primary interest lies in estimating variances, while fixed factors are used for estimating

the mean effect of a treatment (Merlo *et al.* 2005c). Random effects are often used for controlling for correlated structure in the data, that is, dependencies between data ('pseudoreplication'). Nested factors are usually best treated as random effects as we describe below.

In a model with fixed and random factors, it is important to consider how the levels of the fixed factor are related to the levels of the random factor. Their relationship can be nested or crossed (Fig. 3). We will call a fixed factor whose levels vary among groups (of a random effect) a group-level predictor (Gelman & Hill 2007; Kirk 2009; sometimes called 'outer factor'; see Pinheiro & Bates 2000). For example, a treatment might have been applied to randomly selected individuals, and multiple observations were taken per individual. Individuals (a random effect) are nested within treatments, and observations are nested within individuals (and treatments). In this example, 'treatment' is a group-level predictor (outer factor to individual). A fixed factor whose levels vary within groups is called a data- or unit-level predictor (Gelman & Hill 2007; sometimes called 'inner factor' Pinheiro & Bates 2000). For instance, multiple sibships (families are treated as a random effect) might have been split in two treatment groups with one observation per individual. Individuals are clustered within families, but the treatment is crossed to the family random effect. In this instance, 'treatment' is a data-level predictor (inner factor to individual).

In cases of multiple levels ('higher-order hierarchical models'), it might be necessary to be more specific about the different grouping levels. For example, if there are observations clustered in subjects that are nested in families, there are two grouping levels and a statement about a group-level predictor will be ambiguous. In this example, it would be more precise to talk about data-level predictors (the level of observations), subject-level predictors and family-level predictors. If the treatment is applied to whole families, 'treatment' will be a group-level predictor (an outer factor to family). If the treatment is applied to individual subjects, 'treatment' will be a data-level predictor (an inner factor to family, but an outer factor to subject).

Fixed and random factors

Our discussion so far has been rather independent of whether effects are fitted as random or as fixed factors. A full discussion of random and fixed effects goes beyond the scope of this paper, but we will nevertheless point to a few relevant points in the discussion.

Each random effect in a mixed model is modelled as a separate group-level model (Gelman & Hill 2007), in the simplest case by assuming such effects at one level can be modelled as stemming from a common distribution and properties of this distribution are estimated by hyperparameters. In the typical case of normally distributed random effects, the group-level model consists of a normal distribution with a mean of zero and a group-level variance, which is estimated from the data. As with the homoscedasticity of the residuals (cf., Cleasby & Nakagawa 2011), the assumptions about the group-level distributions should also be validated. The assumption of a Gaussian random-effects distribution is likely to be fulfilled if the source of variation is polygenic with small average effect sizes. Strong main effects at the grouping level (such as age classes, sexes, morphs) that are not explicitly modelled will tend to produce bi- or multimodal distributions of group means that are not adequately captured by normal distributions. To put it another way, the levels of the grouping variance should be drawn from a homogeneous population. Outliers (as may be produced by a class of 'others' that include a diverse group of features) will tend to impair the fit of the group-level model and thus the fit of the mixed model as a whole. Whether or not the distribution of random effects follows the assumed distribution can be visually checked using quantile–quantile (Q–Q) plots or histograms, which will help to identify severe violations. We also recommend using common sense, such as reasoning if in the biological system under study it is likely that the sum of small effects will produce approximately normal distributions (which would be ensured by the central limit theorem). If random effects are not normally distributed, it is possible to apply other distributional assumptions on random effects, but again these

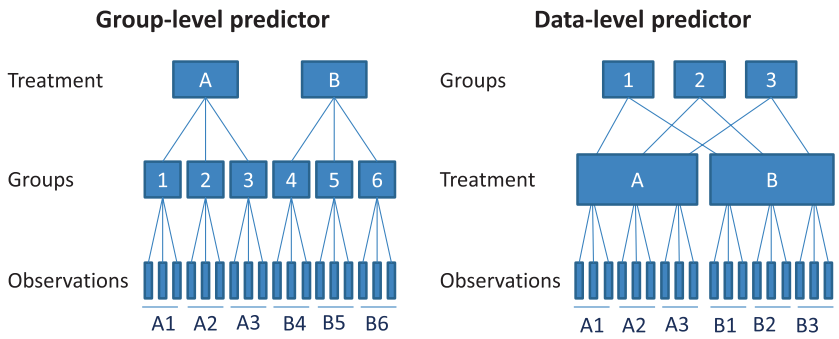

Fig. 3. Group- and data-level predictors in a model with one main effect of interest ('treatment', a predictor that is modelled as a fixed effect) and some clustering in groups (groups are modelled as random effects). With group-level predictors, groups are effectively nested in treatments, whereas with data-level predictors, groups are crossed with treatments. Only with data-level predictors, it is possible to test for interactions, that is, to what degree the effect of the treatment varies among groups. If groups are individuals and the treatment is a drug application, for example, the interaction is caused by differences in the susceptibility of different individuals to the drug. Treatment levels are labelled by upper-case, and groups are labelled by numbers.

assumptions should be validated. In some circumstances, it might also help to fit these 'random factors' as fixed factors instead, which avoids distributional assumptions altogether. In such cases, however, inferences will be limited only on the particular sample of groups.

In practical applications, researchers will usually have their own opinion on whether a particular effect should be fitted as a random or a fixed effect (see Bennington & Thayne 1994 for practical guidelines). The difference certainly matters for model fitting, as random effects are not estimated independently (they are affected by shrinkage), whereas fixed-effect levels of the same predictor are estimated independently of each other. In the context of this paper, it is important that nested effects are more easily fitted as random effects, because this avoids potential pitfalls with choosing the wrong residual variance (Gelman 2005).

Nested factors are typically conceptually random factors (Quinn & Keough 2002). Indeed, it is hard to imagine an experimental design with nested fixed effects, where each level is of particular interest. This is because if we are applying two treatments (each with multiple levels) in a nested manner, the nested treatment is completely confounded with the higher-level treatment effect and it is impossible to estimate meaningful main effects. Although we stated that nested effects are typically (conceptually) random effects, they can still be fitted as fixed effects in a linear model (e.g. Quinn & Keough 2002; Gelman 2005; Kirk 2009). This will be particularly useful if the number of groups is low, which makes it difficult to estimate the group-level variance using mixed models.

Interpreting random-effect variances

Random-effect variances are often not reported and therefore also not interpreted in ecological and evolutionary biology papers. Instead, the interpretation of statistical models is limited to the fixed effects, even if mixed models are fitted. This practice is unfortunate, because the estimators for the random-effect variances allow important biological insight (Merlo *et al.* 2005a,b,c, 2006). Some fields of research such as quantitative genetics (Falconer & Mackay 1996; Lynch & Walsh 1998) or the literature on personality traits and phenotypic plasticity (Dingemanse *et al.* 2009; Martin *et al.* 2011) are indeed focused on the interpretation of variance components. A critical quantity is the between-group variance that can be standardized by the total phenotypic variance to give the repeatability (a form of an intraclass correlation), such that it can be compared across studies (Nakagawa & Schielzeth 2010).

If the random-effect variance is low, there is little potential for strong group-level fixed effects (although they might still become significant with sufficient data, Merlo *et al.* 2005c). In such situations, it might be more fruitful to collect information on data-level predictors in order to explain residual variance and thus for making more accurate predictions. If the random-effect variance is large, however, it might be promising to collect information on group-level predictors because there is great potential that they explain group-level variance. Significant (important) group-level predictors will tend to reduce the random-effect variance, as they explain the part of it and hence reduce the unexplained group-level variance. Unless the random-effect variance is very large, it might still be useful to consider data-level predictors, which can potentially reduce residual variance (Snijders & Bosker 2011; Fig. 3). A reduction of the residual variance is useful even if the main interest is in group-level predictors, as reduced residual variance tends to reduce the standard errors and therefore improves the estimates for group-level predictors. These considerations are most important at the design stage of a follow-up study, although in our experience, variance component analyses based on an examination of random-effect variances can be a useful tool in exploratory data analysis. It is important to consider the potential for type I errors when testing a larger number of data-level predictors (Forstmeier & Schielzeth 2011). A model of general value should not be optimized for a particular data set by extensive 'fishing'.

For example, if we were interested in understanding the survival probability of some species of insect, we might have sampled multiple individuals from multiple patches (patch identity is the random effect). If we found substantial between-patch variance, it would be promising to consider patch-level predictors, such as temperature, exposition, predators, competitors. If we found low between-patch variances, we might want to consider individual-level (i.e. data-level) predictors such as body size, emergence time, mating status. Many data-level predictors such as body size might vary both within and between patches (for genetic or environmental reasons), and it might be useful to separate the two components (van de Pol & Wright 2009; Algina & Swaminathan 2011) certainly if there are some indications of significant between-group variance (e.g. Steinmeyer *et al.* 2010).

Because random-effect variances (or, equivalently, standard deviation of the between-group variation) contain relevant biological information, they should be presented in published papers even if the main aim was to deal with correlated structures ('pseudoreplication') when estimating fixed effects. The random-effect variance along with the residual variance can be presented. A standardized measure of the random-effect variance is the intraclass correlation coefficient. Confidence intervals of variance components are not always provided by standard statistical software, even though such uncertainty estimates would be very valuable for meta-analyses. Confidence intervals for variance components in mixed models can be estimated, for example, by (parametric) bootstrapping (Faraway 2006; Nakagawa & Schielzeth 2010).

Nested designs and crossed syntax

Nested effects can usually be fitted using the syntax for crossed effects if the coding reflects implicit nesting. Statistical software will convert whatever syntax is used into design matrices, and these are blind to nested or crossed data collection. The specifics depend on the software (for a worked example in R see Data S1 in Supporting Information), but we want to highlight that crossed fitting is routinely done when fitting random

effects that are nested within fixed factors in a mixed model. Also, two random factors can be simply fitted as crossed effects if the coding reflects the implicit nesting of the study design. As we have described above, the key difference lies in the interpretation of the model estimates and not in the model fitting itself. We assume that nested factors will typically be fitted as random effects. It is also possible to fit nested ANOVAs with two fixed effects, but this case needs special consideration for the estimation of the standard errors for the higher-level factor that need to be based on the appropriate degrees of freedom (Gelman 2005). We will now discuss the three realistic situations in more detail (there is no meaningful case where a fixed effect is nested within a random effect).

In the case of a random-effect and fixed-effect group-level predictor, crossed fitting is often done without much consideration. Hence, applying a treatment to individuals, taking multiple measurements per individual and fitting treatment as a fixed and individual as a random effect technically fit two crossed effects. The interaction variance is pooled with the random-effect variance. If the treatment is applied to a population of clonal lines, for example, the random-effect variance would capture genetic effects, but also genotype by environment interactions (G × E) (Via & Lande 1985) and is therefore biologically relevant. The genetic variance is inflated by G × E if the data are nested by design.

The two predictors might also both be treated as random effects, for example when analysing a nested half-sib–full-sib breeding design (i.e. North Carolina I). If the factor levels of the nested random effect are labelled uniquely within the whole data set (and the coding thus reflects implicit nesting), the two effects can simply be fitted as two crossed random effects. The interaction variance is pooled with the nested random effect (in this case with the 'dam' variance component). The interaction variance would have a meaningful interpretation, because it includes the dominance interactions between haplotypes.

The two predictors might both be treated as fixed effects. In this case, it is not possible to fit them in a straightforward way as two crossed effects, because such a model would be overparameterized. For example, we might want to fit a model with block and treatment as two factorial predictors where blocks are nested in treatments and both are fitted as fixed effects. We can only fit an intercept, the treatment main effect and an interaction term (treatment × block), but not block main effect (Gelman 2005). The complication when fitting nested fixed effects lies in the fact that the degrees of freedom for estimating the standard error of the higher-level factor (treatment) should be based on the number of levels of the nested factor (block), not on the total number of observations (Gelman 2005). This issue can be avoided by routinely fitting nested factors as random effects.

Extensions and outlook

We have focused on univariate linear models with Gaussian error distributions, because these are most widely used in the fields of ecology and evolution. The mixed model framework, however, is more powerful and allows the explicit modelling of non-Gaussian error distributions. Variance decomposition is somewhat more involved when using generalized linear mixed models (Nakagawa & Schielzeth 2010). Furthermore, mixed models can also be used for fitting nonlinear models, which again demonstrates the generality of the concept. Mixed models can also be used to fit models to multiple responses, so that variances as well as covariances can be estimated on multiple levels (e.g. MCMCglmm package in R, Hadfield 2010). Flexibility also comes with the cost of reduced user-friendliness in cases where no standard cookbook recipe exists. Bayesian approaches are particularly flexible, and the software WinBugs (see Kéry 2010) offers possibilities for fitting linear and nonlinear models with great flexibility in the error as well as in the random-effect distributions. The discussion about the pooling of the variances in nested designs applies regardless of whether they are fitted in a (frequentist) likelihood framework or using a Bayesian approach.

We have focused on blocked random effects in this article. We call them blocked effects, because the design matrix for the random effects consists of 0s and 1s and can be sorted so that the 1s occurs in blocks (Bolker *et al.* 2009; see Fig. 1). Blocked random effects constitute a classic case of clustered data, because each observation is associated with exactly one grouping level. Mixed models can also include random effects that are continuous, for example additive genetic or phylogenetic relatedness matrices (Kruuk 2004; Hadfield & Nakagawa 2010), spatial distance matrices (Valcu & Kempenaers 2010) or multiple membership models (Browne, Goldstein & Rasbash 2001) that relate observations to the grouping level in a more complex fashion.

Several of our examples refer to nesting of observations within individuals. By fitting individual as a random (or fixed) effect, we implicitly assume that observations within individuals are sufficiently independent of each other. If the data collection is designed as a longitudinal study of, for example, growth, then observations closer in time might be more similar to each other than observations taken at greater time intervals (Ives & Zhu 2006). Such data will require fitting of more complex models that control for the temporal structure (Verbeke & Molenberghs 2001; Singer & Willett 2003). Importantly, it is not the nonindependence of the observations that is problematic, but the nonindependence of the residuals. If nonindependence is adequately controlled for by covariates or (blocked or continuous) random effects, a model with interdependent data points might have perfectly uncorrelated residuals. For example, we might find a highly bimodal and thus certainly nonnormal distribution of the raw data. If this non-normality is fully explained by sexual dimorphism or by different age classes and we include the relevant predictors (sex or age, respectively) in the model, then there is no violation of the distributional assumptions.

The difference between nested and crossed effects lies largely in the interpretation. There are three main messages to be extracted. First, a conceptual approach in terms of variance components helps to avoid misinterpretations. In the context of this paper, it is the interaction variance that matters most. Hence, one important question to ask is: Which of my

model estimates will contain the interaction variance? Second, already at the design stage of a study, we should decide if and how we will be able to estimate the interaction variance. A nested design simply leaves fewer 'degrees of freedom' to model all the effects that are potentially of interest. In particular, a nested design is not suited for separating main effect and interaction variance. Third, mixed-effects modelling is a powerful tool for fitting models to structured data. Important biological insight can be gained from evaluating the random-effect variances even if they are not the prime interest of the study.

Acknowledgements

We thank Roger Mundry and Wolfgang Forstmeier for always fruitful discussions and Leif Engqvist, three anonymous referees and members of the stats club Bielefeld for very helpful comments on an earlier version of the manuscript. H.S. was supported by an Emmy-Noether fellowship of the German Research Foundation (SCHI 1188/1-1).

References

Aiken, L.S. & West, S.G. (1991) *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications, Newbury Park, California.

Algina, J. & Swaminathan, H. (2011) Centering in two-level nested designs. *Handbook of Advanced Multilevel Analysis* (eds J.J. Hox & J.K. Roberts), pp. 285–312. Routledge, New York, New York.

Bennington, C.C. & Thayne, W.V. (1994) Use and misuse of mixed-model analysis of variance in ecological studies. *Ecology*, **75**, 717–722.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.-S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, **24**, 127–135.

Browne, W.J., Goldstein, H. & Rasbash, J. (2001) Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, **1**, 103–124.

Cleasby, I.R. & Nakagawa, S. (2011) Neglected biological patterns in the residuals: a behavioural ecologist's guide to co-operating with heteroscedasticity. *Behavioral Ecology and Sociobiology*, **65**, 2361–2372.

Comstock, R.E. & Robinson, H.F. (1952) Estimation of average dominance of genes. *Heterosis* (ed. J.W. Gowen), pp. 494–516. Iowa State College Press, Ames, Iowa.

Congdon, P. (2007) *Bayesian Statistical Modelling*, 2nd edn. Wiley, Chichester.

Diaz-Uriarte, R. (2002) Incorrect analysis of crossover trials in animal behaviour research. *Animal Behaviour*, **63**, 815–822.

Dingemanse, N.J., Kazem, A.J., Reale, D. & Wright, J. (2009) Behavioural reaction norms: animal personality meets individual plasticity. *Trends in Ecology & Evolution*, **25**, 81–89.

Engqvist, L. (2005) The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. *Animal Behaviour*, **70**, 967–971.

Falconer, D.S. & Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*, 4th edn. Prentice Hall, Harlow, UK.

Faraway, J.J. (2006) *Extending the Linear Model with R*. Chapman & Hall/CRC, Boca Raton, Florida.

Forstmeier, W. & Schielzeth, H. (2011) Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology*, **65**, 47–55.

Gelman, A. (2005) Analysis of variance: why it is more important than ever. *Annals of Statistics*, **33**, 1–31.

Gelman, A. & Hill, J. (2007) *Data analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, UK.

Goldstein, H. (2011) *Multilevel Statistical Models*, 4th edn. Wiley, Oxford.

Hadfield, J.D. (2010) MCMC methods for multi-response Generalized Linear Mixed Models: the MCMCglmm R package. *Journal of Statistical Software*, **33**, 1–22.

Hadfield, J.D. & Nakagawa, S. (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, **23**, 494–508.

Hinkelmann, K. & Kempthorne, O. (2008) *Design and Analysis of Experiments. Volume 1: Introduction to Experimental Design*, 2nd edn. Wiley-Blackwell, Chichester, UK.

Ives, A.R. & Zhu, J. (2006) Statistics for correlated data: phylogenies, space, and time. *Ecological Applications*, **16**, 20–32.

Kéry, M. (2010) *Introduction to WinBUGS for Ecologists: A Bayesian Approach to Regression, ANOVA, Mixed Models and Related Analyses*. Academic Press, Amsterdam.

Kirk, R.E. (2009) Experimental design. *The SAGE Handbook of Quantitative Methods in Psychology* (eds R.E. Millsap & A. Maydeu-Olivares), pp. 23–45. Sage Publications, London.

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

Lynch, M. & Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates Inc., Sunderland, Massachusetts.

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.

McCulloch, C.E. & Neuhaus, J.M. (2005) *Generalized Linear Mixed Models*. John Wiley & Sons, Chichester.

Merlo, J., Chaix, B., Yang, M., Lynch, J. & Rastam, L. (2005a) A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon. *Journal of Epidemiology and Community Health*, **59**, 443–449.

Merlo, J., Chaix, B., Yang, M., Lynch, J. & Rastam, L. (2005b) A brief conceptual tutorial on multilevel analysis in social epidemiology: interpreting neighbourhood differences and the effect of neighbourhood characteristics on individual health. *Journal of Epidemiology and Community Health*, **59**, 1022–1028.

Merlo, J., Yang, M., Chaix, B., Lynch, J. & Rastam, L. (2005c) A brief conceptual tutorial on multilevel analysis in social epidemiology: investigating contextual phenomena in different groups of people. *Journal of Epidemiology and Community Health*, **59**, 729–736.

Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., Rastam, L. & Larsen, K. (2006) A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology and Community Health*, **60**, 290–297.

Mousseau, T.A. & Fox, C.W. (1998) *Maternal Effects as Adaptations*. Oxford University Press, Oxford.

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

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

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

Quinn, G.P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge.

Rasch, D., Pilz, J., Verdooren, L.R. & Gebhardt, A. (2011) *Optimal Experimental Design with R*. CRC Press, Boca Raton, Florida.

Ryan, T.P. (2007) *Modern Experimental Design*. John Wiley & Sons, Chichester.

Scheiner, S.M. & Gurevitch, J. (2001) *Design and Analysis of Ecological Experiments*, 2nd edn. Oxford University Press, Oxford.

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

Singer, J.D. & Willett, J.B. (2003) *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, Oxford, UK.

Snijders, T. & Bosker, R. (2011) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd edn. Sage, London.

Steinmeyer, C., Schielzeth, H., Müller, J.C. & Kempenaers, B. (2010) Variation in sleep behaviour in free-living blue tits *Cyanistes caeruleus*: effects of sex, age and environment. *Animal Behaviour*, **80**, 853–864.

Underwood, A.J. (1997) *Experiments in Ecology: Their Logical Design and Interpretation Using Analysis of Variance*. Cambridge University Press, Cambridge.

Valcu, M. & Kempenaers, B. (2010) Spatial autocorrelation: an overlooked concept in behavioral ecology. *Behavioral Ecology*, **21**, 902–905.

Verbeke, G. & Molenberghs, G. (2001) *Linear Mixed Models for Longitudinal Data*. Springer, New York, New York.

Via, S. & Lande, R. (1985) Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution*, **39**, 505–522.

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1**, 3–14.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A. & Smith, G.M. (2009) *Mixed Effects Models and Extensions in Ecology with R*. Springer, London.

Received 22 June 2012; accepted 30 August 2012
Handling Editor: Robert Freckleton

Supporting Information

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

Data S1. Worked example with R code.

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.