Behavioral Ecology
doi:10.1093/beheco/arn145
Advance Access publication 27 November 2008

Conclusions beyond support: overconfident estimates in mixed models

Holger Schielzeth and Wolfgang Forstmeier
Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, PO Box 1564, 82305 Starnberg (Seewiesen), Germany

Mixed-effect models are frequently used to control for the nonindependence of data points, for example, when repeated measures from the same individuals are available. The aim of these models is often to estimate fixed effects and to test their significance. This is usually done by including random intercepts, that is, intercepts that are allowed to vary between individuals. The widespread belief is that this controls for all types of pseudoreplication within individuals. Here we show that this is not the case, if the aim is to estimate effects that vary within individuals and individuals differ in their response to these effects. In these cases, random intercept models give overconfident estimates leading to conclusions that are not supported by the data. By allowing individuals to differ in the slopes of their responses, it is possible to account for the nonindependence of data points that pseudoreplicate slope information. Such random slope models give appropriate standard errors and are easily implemented in standard statistical software. Because random slope models are not always used where they are essential, we suspect that many published findings have too narrow confidence intervals and a substantially inflated type I error rate. Besides reducing type I errors, random slope models have the potential to reduce residual variance by accounting for between-individual variation in slopes, which makes it easier to detect treatment effects that are applied between individuals, hence reducing type II errors as well. *Key words:* experimental design, maternal effects, mixed-effect models, random regression, repeated measures, type I error. *[Behav Ecol 20:416–420 (2009)]*

The development of mixed-model methodology has proceeded rapidly over the past years (Snijders and Bosker 1999; Pinheiro and Bates 2000; Venables and Ripley 2002; Faraway 2006; Galwey 2006; Gelman and Hill 2007), and mixed models are now commonly used in analyses of observational and experimental data. Among other applications, mixed models can be used to control for the nonindependence of data points stemming from the same individual, when the aim is to estimate fixed effects (main effects and/or interactions) and to test their significance. This is often done by including individual-specific intercepts, and the widespread belief is that this controls for all types of pseudoreplication. Here we show that in some widespread test situations, this is not the case. In such situations, random intercept models produce standard errors (SEs) and consequently significance levels that are biased. This is obviously problematic, although point estimates of effect sizes are still unbiased. We show that these problems can be solved by instead using random slope models. This allows drawing conclusions that are actually supported by the data. Furthermore, we call for caution in the assessment of published results that have used inappropriate models.

THE PROBLEM

The problem is most easily illustrated for designs, in which some factor varies between subjects, but other factors (or covariates) vary within subjects (e.g., split-plot and repeated-measures designs; Quinn and Keough 2002). For example, in a study of differential allocation, females are paired experimentally to either attractive or unattractive males (Bolund et al. forthcoming). They are allowed to produce a clutch, and egg sizes are measured for all eggs. When the interest is to estimate the effect of the treatment (attractive vs. unattractive male) on mean egg size, it is sufficient to include individual-specific random intercept effects, that is, allowing females to differ in their mean egg sizes and hence intercepts. This will effectively control for the nonindependence of eggs coming from the same female when the factor of interest, the treatment, is applied to some of the females. However, many studies also focused on how the treatment affects the patterns of female investment over the laying sequence within a clutch. In this case, a model that controls for individual-specific intercepts only, but not for individual-specific slopes of investment (of egg size over the laying sequence), will greatly underestimate the *P* value of 1) the slope main effect and 2) the treatment by laying order interaction, leading to many false-positive findings.

In the above example, the aim is to estimate within-subject slopes (factorial treatments can be considered as slopes as well, Gelman and Hill 2007) and to generalize these slopes to a larger population of individuals from which the subjects were sampled. It is common, however, that individuals do not only differ in their absolute trait value (like mean egg size) but also in their slopes of response to some factor or covariate (like change of egg volume over the laying sequence). By estimating fixed effects, we are usually interested in the average slope in a population of individuals. If there is high between-individual variation in slopes, then taking more measurements from the same individual will make the estimate of this particular slope more precise. However, these additional measurements do not contribute much to make the estimate for the population slope more accurate. Only by measuring more individuals and, hence, more slopes, one can be more confident about the average slope in the population. Problematically, random intercept models wrongly treat repeated measurements within individuals as independent data points with respect to the population slope.

Address correspondence to H. Schielzeth. E-mail: schielz@orn.mpg.de.
Received 26 March 2008; revised 27 October 2008; accepted 27 October 2008.

Hence, estimating slopes from within-individual replicates will give too narrow confidence intervals for the population. In the framework of null hypothesis testing, this will lead to too many rejections of the null hypotheses when testing the population-wide mean slope against some specific value (slope main effect) or the slopes of 2 populations against each other (slope-by-treatment interactions).

The magnitude of the problem depends on 3 factors:

(1) The most critical is the between-individual variation in slopes (Figure 1a). If there is no variation in slopes between individuals, measurements from the same individual can be considered independent from each other and analyzed as if collected from different individuals. In principle, between-individual variation in slopes is independent of between-individual variation in intercepts, but the most common situation is to find both, between-individual variation in slopes and between-individual variation in intercepts, in real data sets.

(2) Within-individual scatter around the individual regression line dilutes the effect of varying slopes (Figure 1b). If there is high within-individual variation, then between-individual differences in slopes might be less important because they explain less of the total variance.

(3) The issue of pseudoreplication increases with the number of measurements taken from the same individual irrespective of how many levels the covariate has (Figure 1c). Already the second measurement from the same individual represents a pseudoreplicate with respect to the population slope. Only estimates derived from single measurements on different individuals are completely independent.

To illustrate the phenomenon of inflated rates of type I error, we generated data sets that mimic data collected from a split-plot design. We randomly assigned 30 virtual individuals to 2 treatments (15 individuals in each treatment). Within individuals, we sampled 5 trait values and the order of these values as a covariate (analogous to egg sizes within a laying sequence of 5 eggs from 1 clutch). We allowed the 30 individuals to vary in their trait value increase over the sequence by drawing slopes from a normal distribution (with a mean of zero and a standard deviation [SD] of σ_b). Furthermore, we allowed within-individual error by assigning single measurements a deviation from the regression slope drawn from a normal distribution (with a mean of zero and an SD of σ_r). There was no population difference in means between treatments, population slopes were zero in both treatments, and there was no between-individual variation in mean trait values. These are not essential assumptions of the simulation because introducing differences between treatments (in slopes and/or intercepts) as well as allowing individuals to vary in their mean trait values gave the same results.

We fitted a random intercept model [lmer (trait~Treatment*LaySeq+(1|IndID))] to the 150 data points using lmer from the lme4 package in R 2.6.2 (Bates 2007; R Development Core Team 2008). For each randomly created data set, we evaluated whether the confidence intervals for the fixed-effect estimates included the true value. In our simulation, the true values were zero so that the proportion of simulations for which the confidence interval did not include zero is the type I error rate. We let the between-individual variation in slopes (σ_b) and within-individual scatter around the regression line (σ_r) vary between 0 and 0.5 and ran 1000 simulations for each parameter combination.

The type I error rate for finding a significant treatment main effect was close to the expected 5% when the between-individual variation in slopes (σ_b) was low, but for high σ_b values, the type I error rate was considerably lower, that is, too conservative (Figure 2, left, panel-wide means: $\alpha = 0.036$ [top] and $\alpha = 0.017$ [bottom]). This reflects a loss of power for testing the between-individual treatment effect when the between-individual variation in slopes was not accounted for. On the contrary, the false-positive rate of finding significant main effects of slopes as well as significant slope-by-treatment interactions increased with the between-individual variation in slopes (σ_b), but this effect got less pronounced as the within-individual scatter around the regression line (σ_r) increased (Figure 2, center and right, panel-wide means: $\alpha = 0.23$ [top center], $\alpha = 0.23$ [top right], $\alpha = 0.095$ [bottom center], and $\alpha = 0.098$ [bottom right]).

To demonstrate how severe this issue can be in a real data set, we used egg size and egg yolk color measurements from 30 zebra finch pairs (Bolund et al. forthcoming). Each female laid 4 clutches, and all 2–6 eggs from each clutch were measured. Zebra finch eggs increased in size over the laying sequence and egg yolks changed from orange toward yellow. We simulated a random assignment of the 30 pairs to 2 fictional treatments separately for the 4 clutches. Because assignment was random, there was, by definition, no true treatment effect and no slope-by-treatment interaction. However, the proportion of

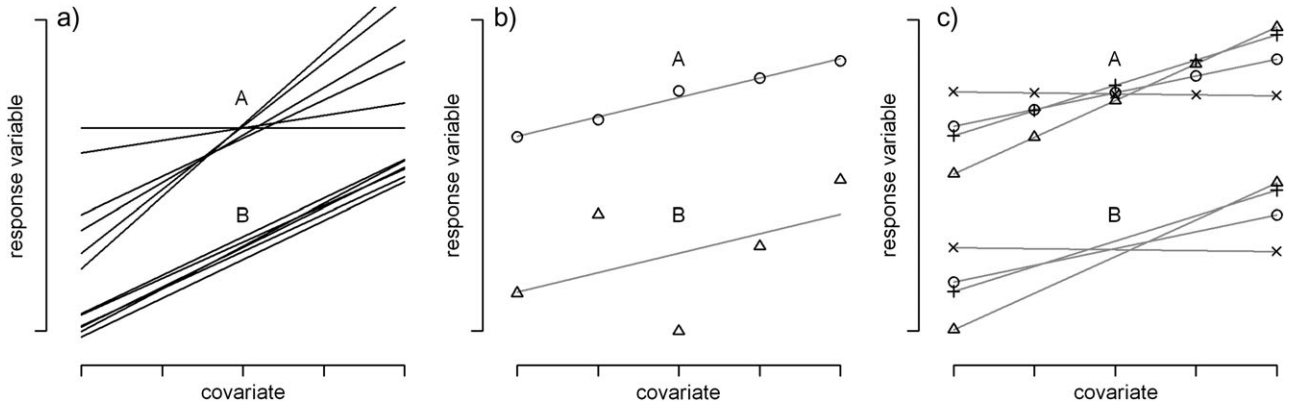

Figure 1
Schematic illustrations of more (A) and less (B) problematic cases for the estimation of fixed-effect covariates in random-intercept models. (a) Regression lines for several individuals with high (A) and low (B) between-individual variation in slopes (σ_b). (b) Two individual regression slopes with low (A) and high (B) scatter around the regression line (σ_r). (c) Regression lines with (A) many and (B) few measurements per individual (independent of the number of levels of the covariate).

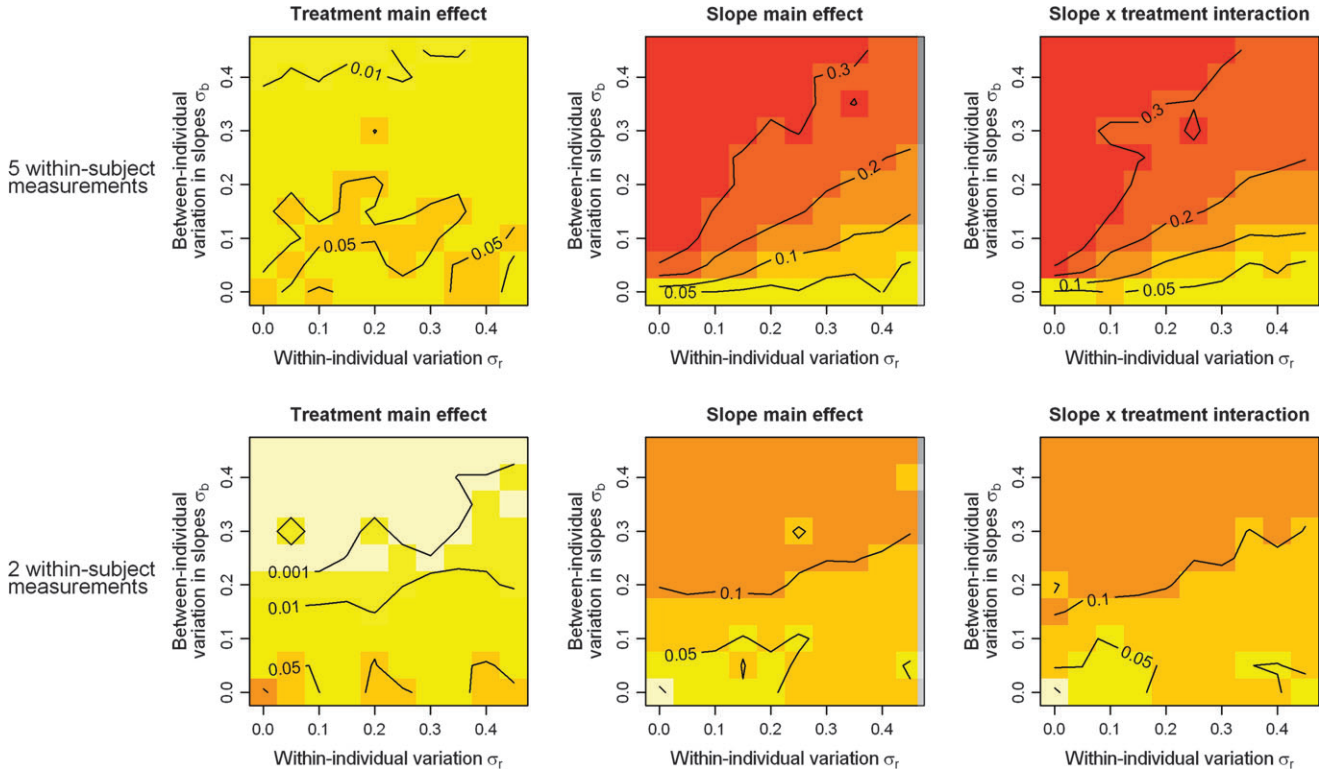

Figure 2
Type I error rate (proportion of estimated 95% confidence intervals for fixed effects not including the true value of zero) in a split-plot design with treatment applied to individuals and slopes measured within individuals as estimated from a random intercept model. The 3 figures in one row show the type I error rates for tests for the treatment main effect (left), the slope main effect (centre), and the slope-by-treatment interaction (right). The 2 rows of figures show simulation with 2 different numbers of measurements within individuals. Type I error rates are indicated by shades of orange (the darker, the higher) and isolines. Error rates depend on the amount of within-individual scatter around the individual regression line (σ_r, *x* axis), and the between-individual variation in regression slopes (σ_b, *y* axis).

significant slope-by-treatment interactions after 10 000 runs ranged from 0.085 to 0.35 (median of 4 clutches: 0.10) for egg size and from 0.11 to 0.41 (median of 4 clutches: 0.15) for egg color. This is clearly more than the desirable rate of false positives ($\alpha = 0.05$).

SOLUTION

The most flexible solution is to use mixed-effect models that include random slopes (and usually also random intercepts) (Laird and Ware 1982; Snijders and Bosker 1999; Raudenbush and Bryk 2002; Singer and Willett 2003;). Random effects are individual-specific effects for intercepts or slopes that are modeled as coming from a common distribution (usually a normal distribution). Hence, unlike linear models that do not include an individual-level model, slopes and/or intercepts are allowed to take different values for each individual. By constraining them to come from a common distribution, individual-level estimates are influenced by measurements from other such individuals (shrinkage to the population mean, Gelman and Hill 2007). If between-individual variation is large, fixed effects are estimated with degrees of freedom close to the number of individuals. If the between-individual variation is low, fixed effects are estimated with degrees of freedom close to the number of data points. This makes sense because in a repeated-measurement design, the evidence for the population slope depends on whether slopes vary mainly within or among individuals. Mixed-effect models will work well for balanced as well as unbalanced data sets because estimates for individual effects (intercepts and slopes) will be weighted by sample size. However, estimates of variance components may become unstable in strongly unbalanced designs (Raudenbush and Bryk 2002).

We analyzed the same randomly generated data set as described in the previous section with mixed-effect models that include random intercepts as well as random slopes [lmer(trait~Treatment*LaySeq+(1 + LaySeq|IndID)]. These models effectively control for pseudoreplication for all predictors by giving the desirable proportion of 5% of the simulations not including the true value (i.e., zero) in the 95% confidence intervals (Figure 3, panel-wide means: $\alpha = 0.048$ [top left], $\alpha = 0.049$ [top center], $\alpha = 0.049$ [top right], $\alpha = 0.049$ [bottom left], $\alpha = 0.048$ [bottom center], and $\alpha = 0.050$ [bottom right]). The random assignment of zebra finch clutches to imaginary treatments yielded a rate of type I error for the treatment by laying order interaction of 0.045–0.054 (median: 0.052) for egg size and 0.053–0.061 for yolk hue (median: 0.053). This shows that random slope models do indeed produce results very close to the desired type I error rate.

Random slope models are easily implemented in standard statistical software (see above for an example of R syntax). They effectively keep the type I error rate at the desired level by giving SEs that are actually supported by the data. The type II error problem in this situation is not greater than normal because only more individuals can confirm that the population slopes differ from some specific value (or between 2 populations that have experienced different treatments). By estimating the between-individual variation in slopes, random slope models make more effective use of the data than, for example, a *t*-test comparing the estimated slopes between 2 treatment

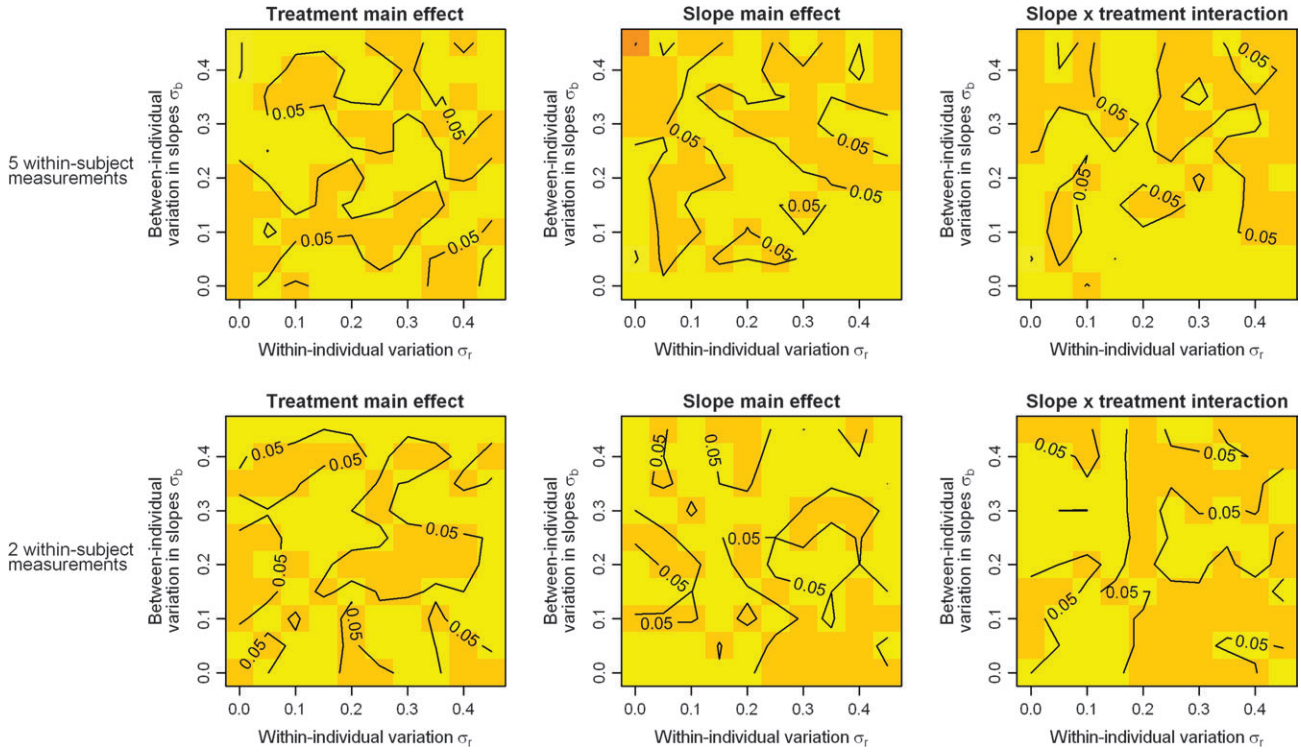

Figure 3
Type I error rates (proportion of estimated 95% confidence intervals for fixed effects not including the true value) in a split-plot design with treatment applied to individuals and slopes measured within individuals as estimated from a random slope model. For details, see Figure 2.

groups. Hence, random slope models are more efficient in detecting effects than the latter. Moreover, they often reduce type II errors for between-individual predictors as compared with random intercept models.

Empirically, some traits will show more between-individual variation in slopes than others (Figure 1a,b). The estimate for the between-individual variance slopes, however, depends on the scaling of both, the covariate and the response. *z*-Transformation of predictor and response (centering to a mean of zero and dividing by the SD) makes this variance better comparable to the residual variance and reduces the correlation between slopes and intercepts (Snijders and Bosker 1999; Raudenbush and Bryk 2002). After *z*-transformation, slopes are estimated as standardized slopes that can vary between −1 and 1 and are equivalent to correlation coefficients (alternatively, Gelman [2008] suggested a standardization by dividing by 2 SDs). Publications should report the variance (or equivalently the SD) of random slopes after *z*-transformation. In our empirical zebra finch example, the random slope SDs are 0.18–0.34 (median 0.23) for egg size and 0.18–0.45 (median 0.24) for yolk hue. The straightforward interpretation is that, in the example of egg volume, slopes of individual females scatter with 0.23 SDs around the populations slope (in this case $b = 0.19$), that is, 95% of all randomly chosen females can be expected to have a slope between 0.64 and −0.26 (estimate ± 1.96 × SD). These large between-individual differences in slopes explain why the type I error rates are greatly inflated (see above).

There are a few potential problems when using random slope models. First, if there are only few individuals, the between-individual variance components are difficult to estimate and tend to be underestimated. This leads to unstable and often slightly overconfident SEs. Second, random slope models might not converge, particularly if more than one random intercept and one random slope are included. The number of parameters to be estimated increases substantially because not only the random effect for the intercepts and slopes but also the correlations among them have to be estimated. In case of convergence problems, we suggest following Figure 1 to judge if including random slopes is likely to have a large influence and to run preliminary submodels to decide whether or not to include particular random slopes.

LITERATURE SURVEY AND TERMINOLOGY

To examine how widespread this problem is, we surveyed empirical papers published between 2004 and 2008 in *Behavioral Ecology*, *Behavioral Ecology and Sociobiology*, *Animal Behaviour* or *Proceedings of the Royal Society of London B* that contained the key words "maternal effects" and "birds." We specifically searched in the avian maternal effect literature because theses studies typically deal with grouped data structures (often multiple eggs/chicks are measured within clutches/broods) and test for within-group predictors (for example laying order). Indeed, 26 of the 37 studies found (70%) analyzed data sets that require accounting for grouping structure. Eighteen of these studies used linear mixed-effect models (69%), 5 studies used repeated-measures ANOVA (14%, 3 of them also used mixed models), 2 used nested ANOVA (8%), 1 used group means to test for group-level predictors (4%), 2 ignored grouping structure (8%), and 1 is unclear about the methods used (4%).

The repeated-measures ANOVA is an appropriate approach, but has a limited applicability, because it only works for balanced data sets. In contrast, nested ANOVA with subjects nested within treatments faces the same problems as a random intercept mixed model. Almost all the 18 studies using mixed-effect models are unclear about the precise nature of their models, but apparently almost all of them have, probably

inappropriately, used random intercept models. Only 2 studies using mixed models explicitly accounted for the issue discussed in this paper: one by using a random intercept model with autoregressive error structure (Grindstaff et al. 2006) and one by using a random slope model (Sockman et al. 2008).

There are several terms used in the statistical literature on mixed models. For example, the terms "hierarchical model" or "multilevel model" are often used as synonyms for the term "mixed-effect model" (Snijders and Bosker 1999; Raudenbush and Bryk 2002; Singer and Willett 2003; Gelman and Hill 2007). This emphasizes that data are modeled on multiple levels: in the examples given in this paper, these are individual level and the data level. However, there might be grouping structures other than (or additional to) grouping by individuals. If there is any ambiguousness, about what constitutes the grouping structure for random effects, this needs explicit specification. In this paper, we refer to random slopes and individual-specific or subject-specific slopes interchangeably. Random slope models are sometimes called random coefficient models (Singer and Willett 2003) or random regression (Schaeffer 2004).

CONCLUSIONS

We have demonstrated that for data sets with grouping structure and within-individual predictors, random slope models are superior to random intercept models both in reducing type II errors for the between-individual predictor (Figure 2, left) as well as reducing type I errors for the within-individual predictors (Figure 2, right). If random intercept models are used inappropriately in such situations, there is a considerable risk of inflated type I errors, depending on how pronounced such random variation in slopes is in any given study system ($0.085 < \alpha < 0.41$ in our example). Probably due to a lack of awareness, published studies do not address this problem adequately in their Methods sections, and therefore, we cannot assess whether analyses have been conducted incorrectly or not. We propose that 1) mixed models should be identified as random intercept or random slope models. Random slope models will usually include random intercepts, too, so that we prefer the shorthand "random slope model" over the more tedious, though more precise, "random intercept, random slope model." This means that an omission of random intercepts in random slope models should be clearly stated. Furthermore, we recommend that 2) fixed factors and covariates, random factors, and interactions between fixed factors and random factors (i.e., random slopes) should be clearly named. Finally, we suggest 3) to report variances or, equivalently, SDs of between-individual variation in slopes on the scale of standardized slopes (i.e., after *z*-transformation of covariates and response). This will help to identify traits where the omission of random slopes is particularly problematic.

FUNDING

The work was funded by an Emmy-Noether fellowship of the Deutsche Forschungsgemeinschaft (FO 340/2 to W.F.).

We thank Stefan van Dongen and Shinichi Nakagawa for important suggestions and critical discussion. Elisabeth Bolund, Henrik Brumm, Alain Jacot, Roger Mundry, and Bart Kempenaers contributed by commenting on early versions of the manuscript. Furthermore, we thank Elisabeth Bolund for providing the empirical data.

REFERENCES

Bates D. 2007. lme4: linear mixed-effects models using S4 classes. R package, 0.99875–9. http://r-forge.r-project.org/projects/lme4/.

Bolund E, Schielzeth H, Forstmeier W. Forthcoming. Compensatory investment in zebra finches: females lay larger eggs when paired to sexually unattractive males. Proc R Soc Lond B Biol Sci. doi: 10.1098/rspb.2008.1251.

Faraway JJ. 2006. Extending the linear model with R. Boca Raton (FL): Chapman & Hall/CRC.

Galwey NW. 2006. Introduction to mixed modelling: beyond regression and analysis of variance. Chichester (UK): John Wiley & Sons.

Gelman A. 2008. Scaling regression inputs by dividing by two standard deviations. Stat Med. 27:2865–2873.

Gelman A, Hill J. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press.

Grindstaff JL, Hasselquist D, Nilsson JA, Sandell M, Smith HG, Stjernman M. 2006. Transgenerational priming of immunity: maternal exposure to a bacterial antigen enhances offspring humoral immunity. Proc R Soc Lond B Biol Sci. 273:2551–2557.

Laird NM, Ware JH. 1982. Random-effects models for longitudinal data. Biometrics. 38:963–974.

Pinheiro JC, Bates D. 2000. Mixed-effects models in S and S-PLUS. New York: Springer.

Quinn GP, Keough MJ. 2002. Experimental design and data analysis for biologists. Cambridge: Cambridge University Press.

R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.

Raudenbush SW, Bryk AS. 2002. Hierarchical linear models: applications and data analysis methods. 2nd ed. London: Sage Publications.

Schaeffer LR. 2004. Application of random regression models in animal breeding. Livest Prod Sci. 86:35–45.

Singer JD, Willett JB. 2003. Applied longitudinal data analysis: modeling change and event occurrence. Oxford: Oxford University Press.

Snijders TAB, Bosker R. 1999. Multilevel analysis: an introduction to basic and advanced multilevel modeling. London: Sage Publications.

Sockman KW, Weiss J, Webster MS, Talbott V, Schwabl H. 2008. Sex-specific effects of yolk-androgens on growth of nestling American kestrels. Behav Ecol Sociobiol. 62:617–625.

Venables WN, Ripley BD. 2002. Modern applied statistics with S. 4th ed. New York: Springer.