RESEARCH ARTICLE

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Robustness of linear mixed-effects models to violations of distributional assumptions

Holger Schielzeth[1] | Niels J. Dingemanse[2] | Shinichi Nakagawa[3] |
David F. Westneat[4] | Hassen Allegue[5] | Céline Teplitsky[6] | Denis Réale[5] |
Ned A. Dochtermann[7] | László Zsolt Garamszegi[8,9] | Yimen G. Araya-Ajoy[10]

[1]Institute of Ecology and Evolution, Friedrich Schiller University, Jena, Germany; [2]Behavioural Ecology, Department of Biology, Ludwig-Maximilians University of Munich, Planegg-Martinsried, Germany; [3]Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW, Australia; [4]Department of Biology, University of Kentucky, Lexington, KY, USA; [5]Département des Sciences Biologiques, Université du Québec à Montréal, Montreal, QC, Canada; [6]Centre d'Ecologie Fonctionelle et Evolutive, CNRS, Montpellier, France; [7]Department of Biological Sciences, North Dakota State University, Fargo, ND, USA; [8]Centre for Ecological Research, Institute of Ecology and Botany, Vácrátót, Hungary; [9]MTA-ELTE, Theoretical Biology and Evolutionary Ecology Research Group, Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös Loránd University, Budapest, Hungary and [10]Centre for Biodiversity Dynamics (CBD), Department of Biology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Correspondence
Holger Schielzeth
Email: holger.schielzeth@uni-jena.de

Funding information
U.S. National Science Foundation, Grant/Award Number: IOS1257718 and IOS1557951; Deutsche Forschungsgemeinschaft, Grant/Award Number: DI 1694/1-1, INST 215/543-1 and 396782608; International Max Planck Research School for Organismal Biology; Centre for Population Biology at the Norwegian University of Science and Technology; Centre d'Ecologie Fonctionelle and Evolutive at the University of Montpellier; Centre for Ecological Research of the Hungarian Academy of Sciences

Handling Editor: Chris Sutherland

Abstract

1. Linear mixed-effects models are powerful tools for analysing complex datasets with repeated or clustered observations, a common data structure in ecology and evolution. Mixed-effects models involve complex fitting procedures and make several assumptions, in particular about the distribution of residual and random effects. Violations of these assumptions are common in real datasets, yet it is not always clear how much these violations matter to accurate and unbiased estimation.

2. Here we address the consequences of violations in distributional assumptions and the impact of missing random effect components on model estimates. In particular, we evaluate the effects of skewed, bimodal and heteroscedastic random effect and residual variances, of missing random effect terms and of correlated fixed effect predictors. We focus on bias and prediction error on estimates of fixed and random effects.

3. Model estimates were usually robust to violations of assumptions, with the exception of slight upward biases in estimates of random effect variance if the generating distribution was bimodal but was modelled by Gaussian error distributions. Further, estimates for (random effect) components that violated distributional assumptions became less precise but remained unbiased. However, this particular problem did not affect other parameters of the model. The same pattern was found for strongly correlated fixed effects, which led to imprecise, but unbiased estimates, with uncertainty estimates reflecting imprecision.

4. Unmodelled sources of random effect variance had predictable effects on variance component estimates. The pattern is best viewed as a cascade of hierarchical

grouping factors. Variances trickle down the hierarchy such that missing higher-level random effect variances pool at lower levels and missing lower-level and crossed random effect variances manifest as residual variance.

5. Overall, our results show remarkable robustness of mixed-effects models that should allow researchers to use mixed-effects models even if the distributional assumptions are objectively violated. However, this does not free researchers from careful evaluation of the model. Estimates that are based on data that show clear violations of key assumptions should be treated with caution because individual datasets might give highly imprecise estimates, even if they will be unbiased on average across datasets.

1 | INTRODUCTION

Biological data often vary on multiple levels and complex data structures have become the norm in the study of ecology and evolution (Allegue et al., 2017; Bolker, 2008; Cheng, Edwards, Maldonado-Molina, Komro, & Muller, 2010; Zuur, Ieno, Walker, Saveliev, & Smith, 2009). Many studies now deal with data that are non-independent at the level of individuals, patches, cohorts or measuring batches. Linear mixed-effects models (LMMs) have become the tool of choice for analysing these types of datasets (Bolker et al., 2009). Unlike standard linear models (LMs), LMMs make assumptions not only about the distribution of residuals, but also about the distribution of random effects (Grilli & Rampichini, 2015). Unfortunately, distributional assumptions for random effects cannot be checked as easily as for fixed effects (Alonso, Litiere, & Laenen, 2010); indeed, most standard software packages do not feature explicit tests of assumptions (though see specialized ʀ-packages like HLM_DIAG, Loy & Hofmann, 2014). However, do such violations affect estimates of parameters of key interest?

Formally, the assumptions of a mixed-effects model involve validity of the model, independence of the data points, linearity of the relationship between predictor and response, absence of measurement error in the predictor, homogeneity of the residuals, independence of the random effects versus covariates (exogeneity), occurrence of data missing completely at random and assumptions about the distribution of the residuals and random effects (Gelman & Hill, 2007; Grilli & Rampichini, 2015; Snijders & Bosker, 2011; Zuur et al., 2009). Another way of stating the core assumptions is that the residuals and random effect coefficients are independent and identically distributed. Any violation of this implies invalidity of the model. Hence, the model needs to be complete in such a sense that all the remaining effects are marginalized by appropriate randomization. Mixed models are flexible, in principle, in their use of various distributions (Lee & Nelder, 2004), but normal distributions are by far the most commonly used. An additional concern is therefore whether violations of the normality assumption affect model fit.

Model diagnosis is typically based on evaluation of the distribution of Cholesky residuals (Cheng et al., 2010). In the ideal case, residuals should be normally distributed—although a normal distribution of residuals does not guarantee that the distributional assumptions are fulfilled. Plotting residuals against fitted values allows an assessment of heteroscedasticity—although again observing the predicted pattern does not guarantee the absence of heteroscedasticity. However, unlike standard linear models, the distributional assumptions in mixed-effects models need to be checked at multiple levels, including the distribution of random effect coefficients (Snijders & Bosker, 2011). Data points with high leverage on model estimates can also be an issue, but such leverage can be assessed with influence diagnosis tools (Demidenko & Stukel, 2005; Loy & Hofmann, 2013; Santos Nobre & Singer, 2007; Zare & Rasekh, 2011).

Distributional assumptions are notoriously difficult to check, particularly for random effects. Since group means are unobservable directly (even more so in generalized linear mixed-effects models), any violation might be due to violations of the random effect distribution or of other parts of the model (Grilli & Rampichini, 2015). Consequently, McCulloch and Neuhaus (2011) suggested checking the distributional assumptions of the lower levels first, before checking the distribution of group levels. A posterior predictive model check offers a general approach for checking the predictive accuracy of a model (Box, 1980; Gelman & Hill, 2007; Rubin, 1984). A posterior predictive model check is based on the simulation of data using parameter estimates of the model (incorporating uncertainty in the estimates) with relevant features for evaluating whether the observed values are within the range of simulated values. The problematic part of this approach is the choice of relevant features for assessment. Ideally, these features should be motivated by biological interest but should not be parameters for which the model is optimized (Sinharay & Stern, 2003). The mean value of a parameter is thus usually a poor choice (because it is estimated in the model), but other aspects of the distributions of observations, such as the occurrence of extreme values, might be.

Various studies have tested the effect of violations of distributional assumptions on model estimates on model estimates, focusing primarily on estimates of fixed effects. Fixed effects have been found to be largely robust to violations of distributional assumptions of the random effects, at least for Gaussian models (Arnau, Bendayan, Blanca, & Bono, 2013; Jacqmin-Gadda, Sibillot, Proust, Molina, & Thiebaut, 2007; Maas & Hox, 2004; McCulloch & Neuhaus, 2011; Sinharay & Stern, 2003; Verbeke & Lesaffre, 1997; Warrington et al., 2014). For generalized linear mixed-effects models, the patterns have been more variable and fixed effect estimates can be biased in some cases (Grilli & Rampichini, 2015; Heagerty & Kurland, 2001). Overall, violations of assumptions regarding random effect distributions appear to have minor consequences for linear models, but potentially have serious consequences for non-linear models, including generalized linear mixed-effects models (Grilli & Rampichini, 2015). Unfortunately, very few studies have assessed random effect variances and predictions (e.g. Maas & Hox, 2004; McCulloch & Neuhaus, 2011; Verbeke & Lesaffre, 1996); for those who did, the estimation of the group variance is usually accurate, but random effect coefficients (best linear unbiased predictors, BLUP) are frequently misestimated.

Here we address several general questions about the robustness of LMMs to violations of distributional assumptions at each level of a model. We focus on parameter estimates, i.e. the slopes for fixed effects and the variance explained by random effects, which are usually of greatest interest to researchers in ecology and evolution. We implement a simulation scheme that features several severe violations of distributional assumptions, including skewed, bimodal and heteroscedastic distributions of residuals, as well as missing random effect components. Such violations of distributional assumptions might arise for a variety of reasons in real datasets, for example if some relevant influence (like the state of an organism or the time of sampling) is not accounted for or if measurements show boundary effects. We independently vary violations of the distributional assumptions of the error variance and the distribution of random effects. We then evaluate the effect on parameter estimates. Overall, we found remarkable robustness of LMMs. Bias is generally small in estimated parameters, with the most pronounced problems arising when predictors or random effect components are missing. A cursory exploration of generalized linear mixed-effects models (GLMMs) shows substantial robustness as well but also some notable complications. Our results do not free researchers from caring about assumptions, but should still encourage the mindful use of LMM even in seemingly problematic cases.

2 | MATERIALS AND METHODS

2.1 | Simulated base model

We simulated data to be fitted in a simple LMM and then purposefully violated assumptions about random effects and error distributions for this model (see Table 1 for definitions of key terms). The base model contained two continuous fixed effects predictors that were, for the sake of simplicity, drawn from uncorrelated normal

TABLE 1 Glossary of key terms

Term	Explanation
Best linear unbiased estimates (BLUE)	Fitted values for specific fixed effect slopes
Best linear unbiased predictors (BLUP)	Fitted values for specific random effect levels (also known as Empirical Bayes estimators or conditional modes)
Bias	Mean difference between the estimated and the true (simulated) value
Fitted values	Model estimates for existing observations
Fixed effects	Factorial or continuous predictors for which the slopes are estimated for each level or covariate without modelling a hyperparameter
Hyperparameter	Unobservable parameter that in the cases covered here typically model the variance among instances
Precision/Prediction error	Square root of the mean squared difference between the estimated and the true value
Predicted values	Model estimates for novel observations
Random effect	Grouping factor for which the variance among levels is estimated by a hyperparameter (that usually is of main interest in the analysis)
Residuals	Residual deviations from fitted values that are unexplained by the model

distributions with zero mean and unit variance. The slope for the dependency of the response variable on each of these covariates (x_1 and x_2) was set to +0.2 for the fixed effect of interest (β_1) and to −0.2 for the second fixed effect (β_2). Since covariates were centered and their slopes were of opposite signs, their expected overall effect on the mean was zero. We used a data generating model based on simulated values for all covariates and thus implicitly assume no measurement error in covariates (though with sampling variance across datasets).

We generated 120 observations per iteration. Observations were clustered in 30 groups with four observations on average per group. Groups may represent individuals sampled multiple times or any other hierarchical structure in the data that results in non-independence. The number of replicates per group varied while ensuring that each group was represented at least once. This simulates unbalanced sampling as it is common in ecology and evolution. We drew group-level means from a normal distribution with a mean of zero and a variance of 0.5 and residual deviations were also drawn from a normal distribution with zero mean on a variance of 0.5. The base data generating model was:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

$$\alpha \sim N(0, \sigma_\alpha^2),$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2),$$

where **y** is a vector of the simulated response, β_0 is the intercept that was set to 1, $\mathbf{x}_1$ and $\mathbf{x}_2$ are the vectors of the covariate values, β_1 and β_2

are the two slopes that were set to $\beta_1 = 0.2$ and $\beta_2 = -0.2$, respectively, α represents a vector of the random effect for group identity, ε vector of the residual errors, σ_α^2 the among-group variance (set to 0.5) and σ_ε^2 the error variance (within-group variance, set to 0.5). The expected phenotypic variance was $\sigma_P^2 = \sigma_\alpha^2 + \sigma_\varepsilon^2 + \beta_1^2\sigma_{x_1}^2 + \beta_2^2\sigma_{x_2}^2 = 1.08$ (Allegue et al., 2017), the expected unadjusted repeatability was thus $\sigma_\alpha^2/\sigma_P^2 = 0.46$, and the expected adjusted repeatability was thus $R_{\text{adj}} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2} = 0.5$ (Nakagawa & Schielzeth, 2010).

2.2 | Benchmark models

We generated 10,000 simulated datasets per scenario (the base model introduced above and the scenarios with violations as introduced below). In all cases we then fitted the analysis model:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \hat{\alpha} + \hat{\varepsilon},$$

$$\hat{\alpha} \sim N(0, \hat{\sigma}_\alpha^2),$$

$$\hat{\varepsilon} \sim N(0, \hat{\sigma}_\varepsilon^2),$$

to the data, where b_0, b_1, b_2, $\hat{\alpha}$, $\hat{\varepsilon}$, $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\varepsilon^2$ are the estimates of β_0, β_1, β_2, α, ε, σ_α^2 and σ_ε^2 respectively. We purposefully mis-specified models, either in the distributional assumptions, by ignoring additional grouping structure or correlations among predictors (see below for details). We evaluated the estimated slope b_1 for the parameters β_1, $\hat{\sigma}_\alpha^2$ for the parameter σ_α^2 and $\hat{\sigma}_\varepsilon^2$ for the parameter σ_ε^2. We also inspected the effect on the standard error of the fixed effects. We assessed bias as the mean deviation from the simulated value and prediction error as the square root of the mean squared deviations of the estimated from the simulated value (root mean square error, RMSE). Our analysis is focused on estimation and less on inference (e.g. type I and type II) errors, since we believe that effect size estimation is most relevant in the long run. However, since we realize that significance testing is important, we also explore the coverage in the confidence intervals for fixed effects.

All simulations were programmed in R 3.6.1 (R Core Team, 2019) and resulting data analysed with the LME4 package (version 1.1-21) for fitting mixed-effects models (Bates, Mächler, Bolker, & Walker, 2015).

2.3 | Violations of distributional assumptions

Analysis of data with LMMs would typically assume that α and ε are sampled from normal distributions as in our base simulations above. To assess if violations of this assumption are a problem, we purposefully violated this assumption by simulating datasets in which we sampled α and ε from other distributions. This was done independently for both terms (group variance and residuals) one at a time and for both in combination.

Skewed distributions (scenario set A): Instead of drawing from standard normal distributions, α and ε were drawn from skewed distributions using the function rpearson from the package PearsonDS (version 1.1, Becker & Klößner, 2017) in R 3.6.1 (R Core Team, 2019) with the skew parameter set to 3. A skew parameter of 0 yields a standard normal distribution and a skew parameter of 3 ensures that skew was recognizably extreme (see Section 3), probably more than in most real scenarios, in order to discover even small biases. The skew parameter (as simulated by the rpearson function) does not change the simulated variance and the expected repeatabilities and phenotypic variances thus remain the same.

Bimodal distributions (scenario set B): Instead of drawing from standard normal distributions with a mean of zero for all observations, group level and/or residual variances were drawn from two distinct normal distributions with the means shifted up and down by ±1.5 units. Such shift might arise if there is an influential fixed effect missing from the model. Shifting was done at random, such that approximately (but not exactly) half of the draws were shifted down and half up. This yielded bimodal distributions with modes separated by 3 units (equivalent to 3 *SD*s) and an expected mean that remained at zero. Separation by 3 *SD*s is sufficiently extreme (see Section 3), probably more than in most real scenarios, in order to discover even small biases. The addition of shifted modes affects the total simulated residual variance. In order to adjust the group level and the residual variance to the target value, we restored the designated variance by $\alpha = \alpha' \frac{\sqrt{\sigma_\alpha^2}}{SD(\alpha')}$ and $\varepsilon = \varepsilon' \frac{\sqrt{\sigma_\varepsilon^2}}{SD(\varepsilon')}$, respectively, where α' and ε', are the group-level deviations and the residual error after adding the shift in means, $SD(\alpha')$ and $SD(\varepsilon')$ are the standard deviations of α' and ε', respectively and σ_α^2 and σ_ε^2 are the target residual and group variances (set to 0.5 as above). However, since this would result in identical residual variances across all simulations (with no sampling variance), we let $\sqrt{\sigma_\alpha^2}$ and $\sqrt{\sigma_\varepsilon^2}$ vary slightly by an amount that sampling of 120 observations from 30 groups would yield when sampled from normal distributions. See Supporting Information for more explanations of how and why the variances were standardized with variability.

Heteroskedastic distributions (scenario set C): Instead of drawing from standard normal distributions, α and ε were drawn from distributions where the variance depended on one of the covariates (x_1) as $\sigma_{\text{Het}}^2 = \sigma^2 + \lambda \cdot (x_1 - \min(x_1))$ where the heteroscedasticity factor λ was set either to 2, 4 or 8. A heteroscedasticity factor of 0 yields a standard normal distribution and a larger heteroscedasticity factor of 8 ensures that heteroscedasticity is recognizably extreme, probably more than in most real applications, to discover even small biases. The incorporation of a heteroscedasticity parameter affects the total simulated residual variance. In order to rescale the residual variance to the target value, we restored the designated variance by $\varepsilon = \varepsilon' \frac{\sqrt{\sigma_\varepsilon^2}}{SD(\varepsilon')}$, where ε', is the residual error after adding the heteroscedasticity factor, $SD(\varepsilon')$ is the standard deviation of ε' and σ_ε^2 is the target residual variance (set to 0.5 as above). However, since this would result in identical residual variances across all simulations (with no sampling variance), we let $\sqrt{\sigma_\varepsilon^2}$ vary slightly by an amount that sampling of 120 observations would yield when sampled from normal distributions.

2.4 | Missing random effects

In order to investigate the effect of missing random effects, we introduced a higher-level random effect (with ten levels), a lower-level random effect (with 60 levels) or a crossed random effect (with 30 levels; **scenario set D**). Additional random effects were sampled randomly, i.e. in an unbalanced fashion, ensuring that each group was represented at least once. The random effect variances were set to $\sigma_H^2 = 0.5$ for the higher-level random effect, $\sigma_L^2 = 0.5$ for the lower-level random effect and $\sigma_C^2 = 0.5$ for the crossed random effect, while the variance of the random effect of interest and the residual variance were kept at $\sigma_\alpha^2 = 0.5$ and $\sigma_\varepsilon^2 = 0.5$ respectively. All additional random effects were normally distributed and thus did not violate any distributional assumptions. Adding additional sources of variance affects the expected repeatability, but this is unproblematic, since we used these simulations primarily to illustrate where, i.e. in which component of variance, the additional unmodelled variance would appear in the model.

2.5 | Correlated predictors

The base version of the model generated covariate values independently (expected correlation of zero). We also introduced cases of correlated predictors, by drawing values of the covariates x_1 and x_2 from multivariate normal distributions with correlations set to +0.2, +0.5 or +0.8 (**scenario set E**). These simulations were introduced to study biases and prediction errors in a situation typically assumed to be problematic. The three correlation values were chosen to show the effect of mild, moderate and strong correlations among predictors. At least the situation with a correlation of 0.8 will typically be considered as very problematic.

2.6 | Further exploration

Besides the models introduced above, we also implemented the following simulations, the results of which are presented as supplementary documents:

1. Balanced sampling: We repeated the entire analysis with a sampling design that was balanced with respect to the distribution of group levels across observations.
2. Small sample data: We repeated the entire analysis with a sample size reduced to 40 observations, 10 groups, five higher-level groups, 20 lower-level groups and 10 crossed-level groups.
3. Few groups: We repeated the entire analysis with a different distribution of observations across groups: 120 observations, six groups, three higher-level groups, 12 lower-level groups and six crossed-level groups.
4. Low repeatability: To evaluate the effect of low repeatability, we also simulated the cases of $\sigma_\alpha^2 = 0.1$ and $\sigma_\varepsilon^2 = 0.9$ (expected adjusted $R = 0.1$) with the rest of the settings as in the base simulation.
5. High repeatability: To evaluate the effect of high repeatability, we also simulated the cases of $\sigma_\alpha^2 = 0.9$ and $\sigma_\varepsilon^2 = 0.1$ (expected adjusted $R = 0.9$) with the rest of the settings as in the base simulation.
6. Generalized LMMs with Poisson data: We repeated the entire analysis with simulated count data. Latent values were simulated as for the Gaussian approach, but observations were generated as samples from Poisson distributions after exponentiating latent values. In order to generate realistic data ranges, the latent scale values were divided by 2 before simulating observations. The data were analysed with log link using the glmer function in R (package LME4, Bates et al., 2015), including an observation-level random effect to model residual variance. We analysed parameter estimates on the latent scale. To make data comparable to other scenarios, we multiplied estimated variances by 4 and slopes by 2, which compensates for the reduced variation on the latent scale.
7. Generalized LMMs with proportional data: We repeated the entire analysis with proportion data. Latent values were simulated as for the Gaussian approach except for the intercept that was set to zero in order to achieve maximum power. Observations were generated as samples from Binomial distributions (after inverse logit transformation of latent values) with 20 trials per observation. The data were analysed with logit link using the glmer function in R (package lme4, Bates et al., 2015), including an observation-level random effect to model overdispersion. We analysed parameter estimates on the latent scale.

3 | RESULTS

Violations of distributional assumptions on either random effect variances or residual variances had surprisingly little biasing effect on the estimates of interest. The only notable exception was bias in the estimate of the group variance when the underlying distribution was bimodal, which resulted in slight upward bias (Figure 4). There were, however, effects on the precision of estimates. Less precise estimates resulted from severely skewed distributions (Figure 1), bimodal distributions (Figure S1) and heteroskedastic residuals (Figure S2). With respect to fixed effects, the increased imprecision is appropriately reflected in increased uncertainty estimates, except for heterogeneous residual variance, in which coverage of the CI was low (Figures S11 and S12).

An interesting pattern emerged in the distributions of group means and residuals when estimated from models that assumed normal distributions (Figure 2): The distribution of best linear unbiased predictors (BLUPs) approximates a normal distribution at low repeatability, while they were clearly non-normally distributed at higher repeatabilities (Figure 2). The reversed pattern appeared for residuals that were clearly non-normal at low repeatabilities and approached normality at high repeatability (Figure 2). These two results illustrate that the estimated distributions approached the distribution assumed by the model if there was little signal for the focal level, while they traced the simulated distribution if the signal in the
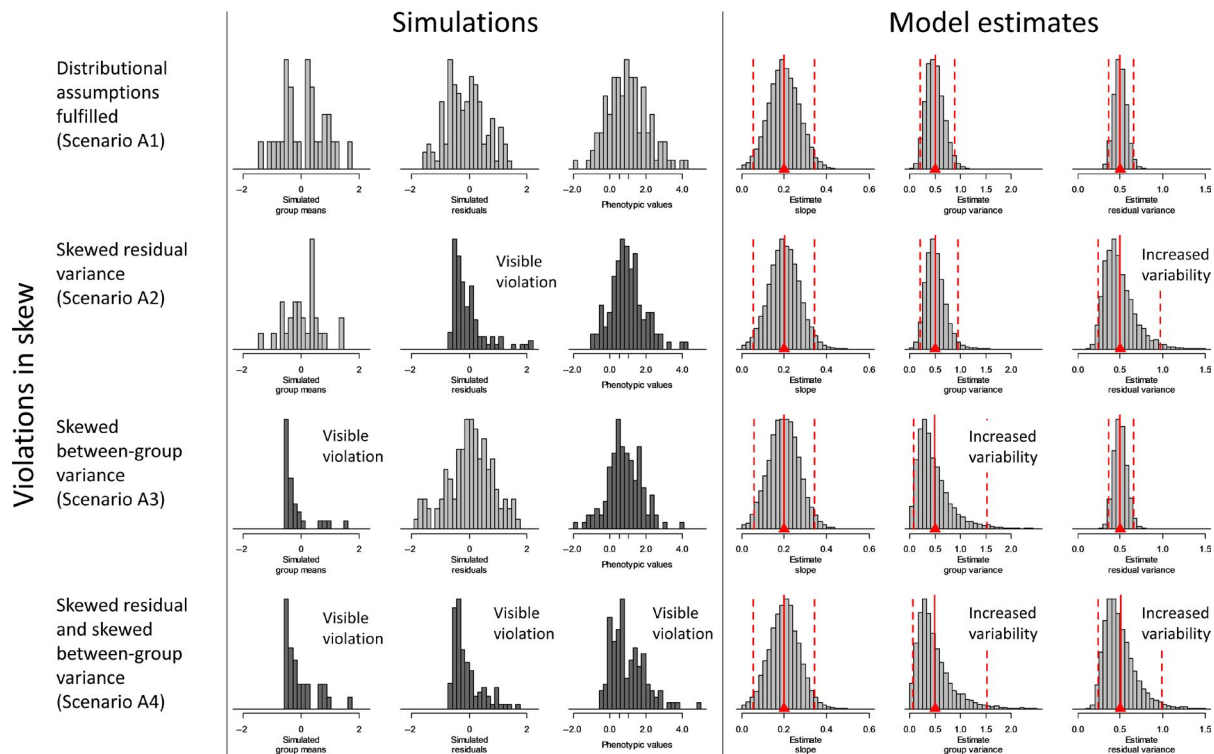
FIGURE 1 Effects of violations of distributional assumptions on parameter estimates of key interest. The left three columns show the distributions of group, residual and phenotypic variance as they were generated in one simulation run. Components that feature violated assumptions are highlighted in dark grey. The three columns on the right show the distribution of point estimates across 10,000 replicated simulations runs for the fixed effect slope, the group variance and the residual variance. The simulated true value is shown as a red triangle. The mean of the estimate values is shown as a solid red line, and the 5% and 95% quantiles of the estimate values are shown as dashed red lines. Upper row: The base mixed model with normal group and residual variances. Second row: Skewed distribution of the residual variances. Third row: Skewed distribution of the group variances. Lowest row: Skewed distributions for the group and residual variances
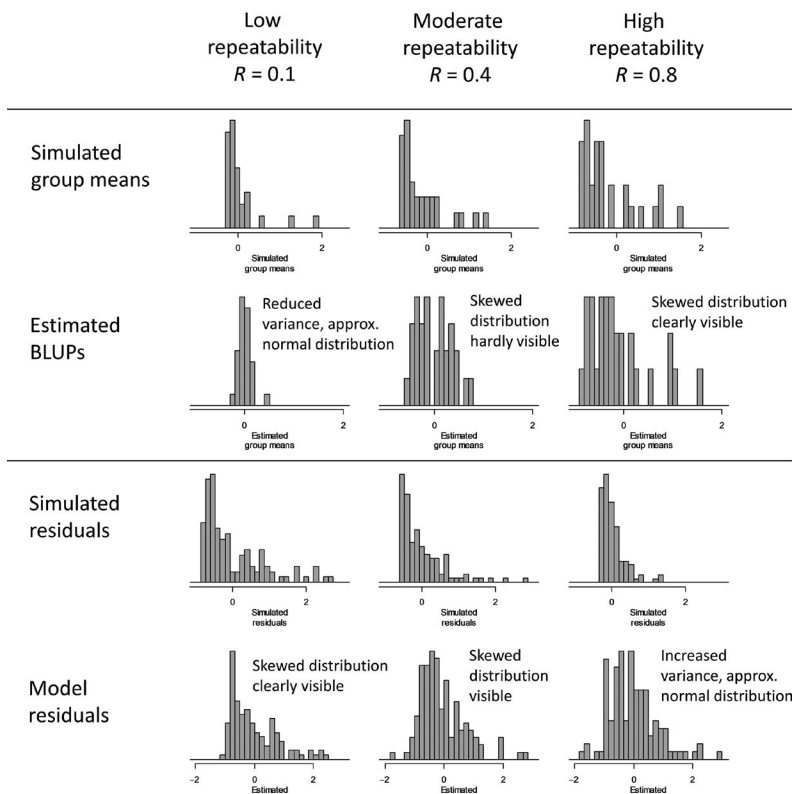
FIGURE 2 Effects of skewed distributions and size of the variance components on estimated group means and residuals. Each column shows a single simulation run for a case with low repeatability (left column), moderate repeatability (middle column) and high repeatability (right column). The upper row shows the distribution of group means as they were simulated from a skewed distribution. The second row shows the estimated group means (BLUPs). The third row shows the distribution of residuals as they were simulated from a skewed distribution. The last row shows residuals as they were estimated by the model
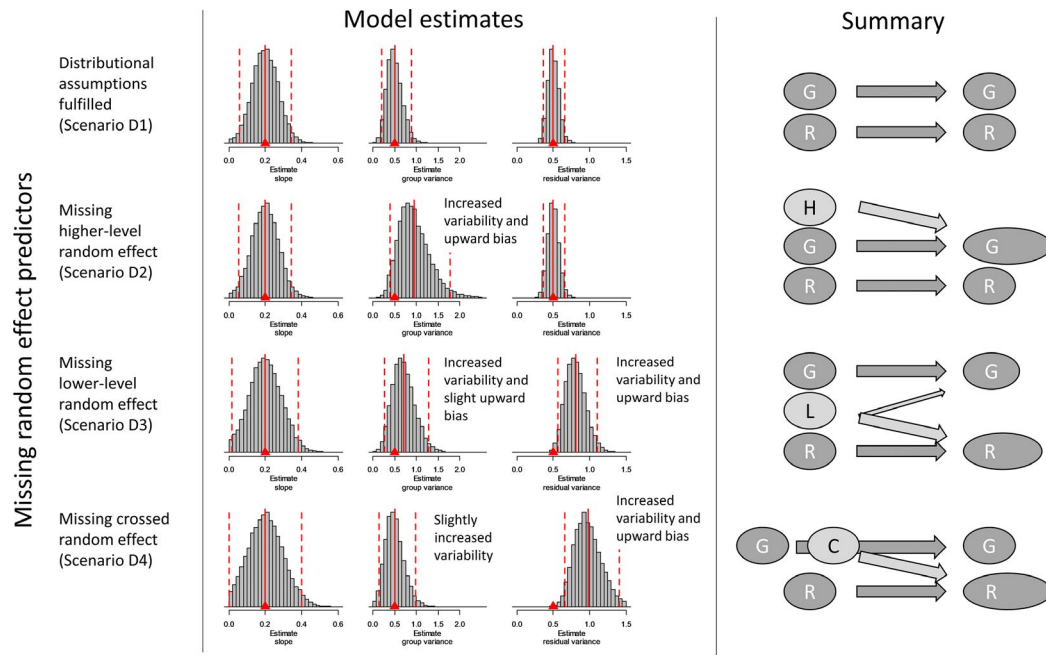
FIGURE 3 Effects of missing random effects on parameter estimates of interest. The three columns on the left show the distribution of point estimates across 10,000 replicated simulation runs for the fixed effect slope, the group variance, and the residual variance. The simulated true value is shown as a red triangle. The mean of the estimate values is shown as a solid red line, and the 5% and 95% quantiles of the estimate values are shown as dashed red lines. The figures on the right illustrate the flow of unmodelled variance components (G, grouping factor of interest; R, residual variance; H, higher-level missing random effect; L, lower-level missing random effect; C, crossed missing random effect). Upper row: The base mixed model with normal group and residual variances. Second row: Missing higher-level random effect with variance of 0.5. Third row: Missing lower-level random effect with variance of 0.5. Lowest row: Missing crossed random effect with variance of 0.5

data was strong, i.e. when the respective variance component was relatively high compared to other variance components.

Missing random effect predictors had little effect on the fixed effect estimates but had systematic effects on the estimates of random effects (Figure 3). The variance due to unmodelled higher-level predictors was almost completely absorbed by the nested random effect variance of interest. The variance arising from unmodelled lower-level predictors was largely absorbed by the residual variance with only a small fraction appearing in the random effect of interest. Even more so, the variance arising from unmodelled crossed random effect predictors was almost completely absorbed by the residual variance.

A correlation between fixed effect estimates resulted in no bias in estimates on average (Figure S3). However, the estimates were less precise and scattered more widely when predictors were correlated, but the effect was marked only for very strong correlations ($r = 0.8$, but not $r = 0.5$ in our simulation). Weak correlations had almost no effect on parameter estimates.

Overall, bias in fixed effect estimates was small in all scenarios, typically <1% of the 'true' value in our simulations (Figure 4). Bias in group effect variances was also <2% in most cases, except for increased upward bias (up to 10%) in cases with bimodal group variance distributions and more extreme deviations in cases of missing random effect predictors as described above. Added prediction error (as compared to the control scenario) was rather small

compared to the control scenarios for fixed effects with minor increase in cases of heteroscedasticity (Figure 4). Random effect components that violated assumptions showed increased prediction error (Figure 4).

Balanced sampling showed few differences from the base scenario with marginally reduced prediction error in the group variance (Figure S4). Simulations with low sample size generally showed a similar pattern, although, as expected, prediction errors were significantly increased for all components (Figure S5). Simulations with few groups show largely similar patterns, but larger prediction error for group variances and increased susceptibility in the case of bimodal group variances (Figures S6).

Simulations with low repeatability did not differ much for estimates of fixed effects, though prediction error was slightly increased, reflecting the overall lower amount of variance explained by the model (Figure S7). Interestingly, the upward bias of the group variance in scenarios with bimodal group distributions was absent at low repeatabilities. Relative bias in the group variance estimate was larger and uniformly positive with significantly larger relative prediction error. Simulations with high repeatability showed the opposite pattern of reduced bias and prediction error in fixed effect estimates and group-level variances (Figure S8).

In simulations with Poisson and binomial distributions, GLMM results also showed generally similar patterns with larger prediction error, reflecting the overall lower information content per data
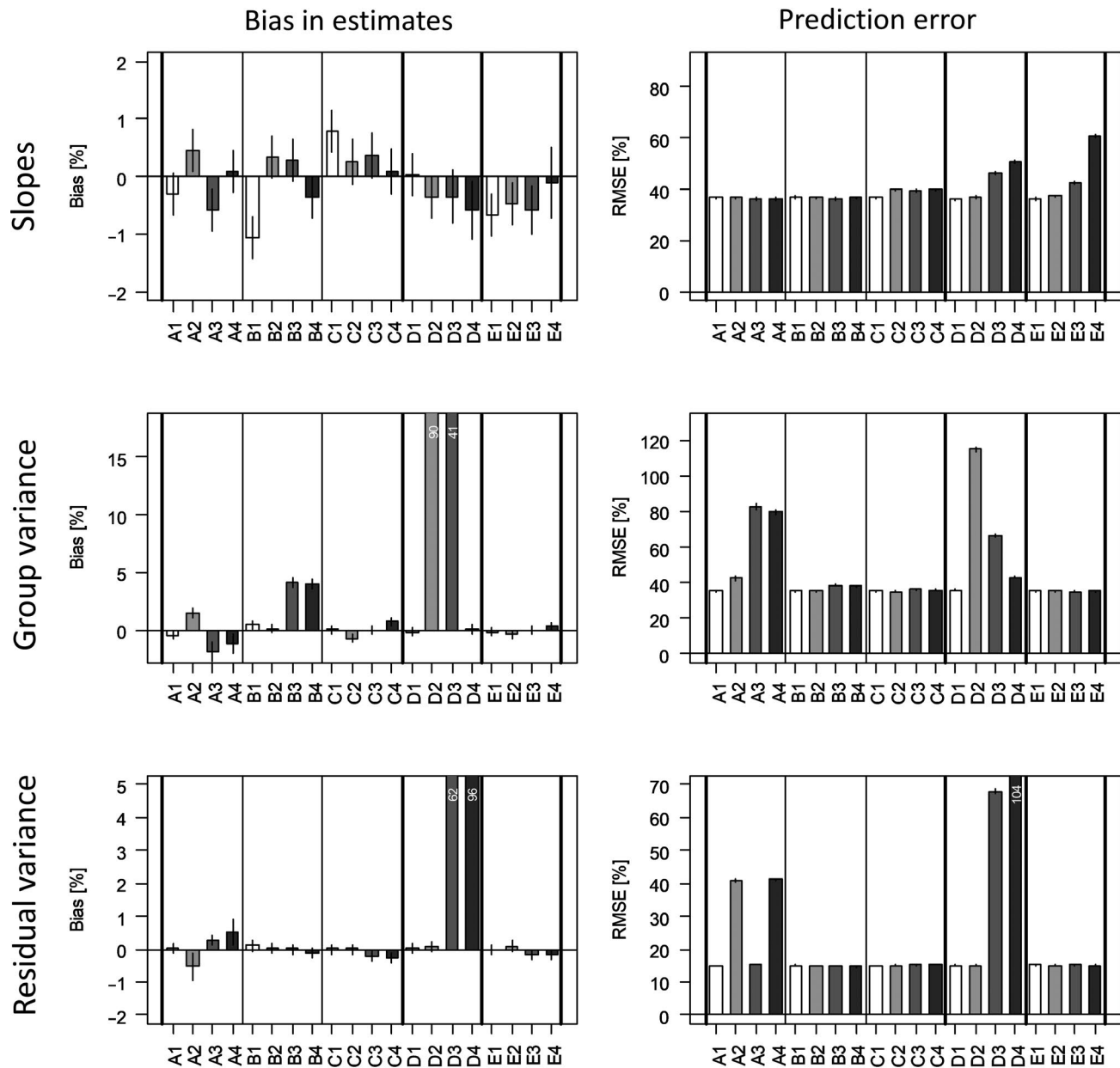
FIGURE 4 Bias and prediction error in fixed effect (upper row), random effect variance (middle row), and residuals variances (lower row) estimates. Vertical lines separate the five different sets of scenarios (A–E) of violated modal assumptions. In each plot, the leftmost bars (A1, B1, C1, D1, E1) refer to the control scenario while the model violations tend to increase from left to right within blocks. Both bias and root mean square prediction error are expressed in relation to the true mean size of the component. Error bars represent the estimated standard error based on 10,000 simulations. Some bars are way beyond the range of the plot and their values are shown as numbers in the top part of the respective bar

point, and consistent downward bias in the group variance even for the control scenario (Figures S9 and S10). However, there were also important differences: Skewed distributions had a significantly larger effect on the respective variance components, resulting in stronger bias in the estimates. Heteroscedasticity affected the

estimate of the residual variance. Unlike the Gaussian scenarios, the fixed effect estimates were also biased in cases of heteroscedasticity (Figures S9 and S10). Coverage of the confidence intervals of fixed effects were too low in particular for Poisson models (Figure S12).

4 | DISCUSSION

Our simulation analysis shows that the effect of violations of distributional assumptions of random effect variances and residuals is surprisingly small. Even substantially skewed, bimodal and heteroscedastic distributions resulted in little overall bias—with the exception of group variances estimated for data generated from a bimodal distribution. Fixed effect estimates in particular were relatively unbiased, though coverage of the confidence intervals was somewhat too low in the case of heterogeneous residual variances. However, the variance components for which the normality assumption was violated showed significantly increased prediction error, meaning that estimates were more variable and thus less precise. Hence, while estimates are unbiased on average, they might be further from the true value than when the distributional assumptions are not fulfilled.

Some of the violations that we simulated could easily emerge from incomplete models. Bimodal distributions, for example, may often arise from a strong effect of a class variable that was not included in the model. Indeed, this was how we generated bimodal distributions in the simulation. Such missing effects might be sex or age classes for the group means or different contexts or environments for the residuals. Identifying and fitting those factors would stabilize the distributions and therefore improve estimates. Similar effects may underlie cases of heteroscedasticity. Skewed distributions frequently arise from effects being multiplicative rather than additive, thus the scale of the measurement should be assessed in such cases (Houle, Pelabon, Wagner, & Hansen, 2011) and appropriate transformation might reduce heteroscedasticity.

We also show that missing random effect terms have systematic effects on estimates of other variance components. Unmodelled sources of variance have predictable effects on model variance components that depend on the hierarchical structure of the dataset. The effects are analogous to a cascade, where variances flow down the hierarchy, but usually not up. Hence, variances of missing higher-level structure cascade down to a lower-level random effect that are included in the model. Lower-level and equal-level (crossed) random effects will cascade down to the residuals. This also implies that when faced with the problem of an incentive to omit one of two hierarchical random effects, it should usually be the lower-level random effect that should be modelled and the higher-level random effect that is left out. This issue might appear in very complex models or if two levels are nearly identical in their factor levels. Modelling the random effect with more levels is then preferable (though the decision also depends on fixed effects, such as group-level predictors). Evidently, this will create some dependencies among some random effect levels, but seems to be not very problematic in the cases that we simulated. Note, however, that we simulated only data-level predictors and the pattern might differ for group-level or mixed-level predictors.

Finally, we show increased imprecision in fixed effect estimates when covariates are correlated. Warnings about correlated fixed effects are abound (Freckleton, 2011; Quinn & Keough, 2002; Zuur, Ieno, & Elphick, 2010), but the impact of such collinearity is much more subtle than commonly assumed. Briefly, large effects on precision of estimates are evident only when fairly strong correlations exist, while weak and moderate correlations have very little effect. Furthermore, while the correlation among predictors affects the estimates of those variables involved in these correlations, the problem does not trickle through to the rest of the model, i.e. it does not affect estimates of random effect components or other uncorrelated predictors. Morrissey and Ruxton (2018) also point out that the increased variation in estimates is a simple consequence of the sampling design being inadequate for estimating independent effects of correlated variables.

A pertinent question when estimates are highly variable is if the uncertainty is appropriately reflected in the estimated standard error of the estimates. Besides the slight under-coverage of confidence intervals caused by heterogeneous residuals variances, fixed effect estimates did not show high levels of imprecision in most scenarios, hence the main concern is the uncertainty estimate for the random effect variance. However, there is no universally accepted way in which uncertainty in random effect variances is quantified. Profile likelihoods, Bayesian estimation or parametric bootstrapping offer three very different options (Gelman & Hill, 2007; Venzon & Moolgavkar, 1988). Our simulation results give some indication for how parametric bootstrapping is likely to perform. Parametric bootstrapping relies on simulations based on model estimates followed by refitting the model and assessing the variability in estimates across simulations. Although we did not perform this for computational reasons (many bootstrap iterations are required per simulation), our results suggest that bootstrapping would not cover the correct estimates well. In cases where the best estimate is far from the true value, parametric bootstrapping will generate data that are more concentrated around the estimated value. Furthermore, parametric bootstrapping will not reproduce the underlying distributional violations that caused the initial outlier estimate. All methods will probably fail in accurately predicting specific random effect levels, as our simulations show that BLUPs were strongly biased towards the assumed distribution if the respective variance component is small (see also Verbeke & Lesaffre, 1996).

A common practice is to perform nonlinear transformations of the response variable in order to improve the fit to normal distributions (Gurka, Edwards, Muller, & Kupper, 2006). However, this comes at a cost of reduced interpretability, when the data are not on the original scale (Grilli & Rampichini, 2015; Houle et al., 2011; Jacqmin-Gadda et al., 2007). Our results suggest that transformations might not be necessary, since violations of the normality assumption often have little impact. It might thus be beneficial to avoid transformation when faced with the trade-off between interpretability and conformance to model assumptions. Furthermore, the decision to transform should consider measurement theory, since not all transformations are consistent with all kinds of data (Houle et al., 2011). One example where transformations should be avoided is in the estimation of selection gradients (Brodie, Moore, & Janzen, 1995). There is a strong theoretical motivation for estimating selection gradients as the slope of relative fitness (absolute fitness divided by average fitness in the population) over trait values. The response, relative

fitness, is almost certainly not normally distributed. Our simulations suggest that the estimate of the relevant slope, the selection gradient, will likely be unbiased in such situations.

We have simulated unbalanced data sampling as it frequently occurs in the study of ecology and evolution. Our results show remarkable robustness even with such unbalanced data. However, our data were missing completely at random (MCAR) with respect to the parameters and response of interest. MCAR is often not the mode of missingness in the study of ecology and evolution. If missing data depend on covariate values they are called missing at random (MAR) while when missing data depend on themselves (i.e. the response variable) or unobserved predictors, they are called missing not at random (MNAR; Nakagawa, 2015; Nakagawa & Freckleton, 2008). This scenario includes the random effect level, when their representation depends on their phenotype of interest (MNAR) or on other observed features (MAR) or unobserved features (MNAR). While missing data imputation techniques can handle MAR, MNAR data are problematic and usually lead to biases whether or not distributional assumptions of the model are fulfilled.

There are approaches for obtaining robust estimators for the fixed effects for cases of small sample size and for models with violated distributional assumptions (e.g. Kenward & Roger, 1997; Kenward & Roger, 2009; Royall, 1986; Royall & Tsou, 2003). Sandwich estimators, for example, increase the standard errors to a degree that small samples or violations would increase uncertainty beyond what is expected from large samples with no violations. Sandwich estimators have repeatedly been shown to perform well, except when there is a combination of violations and very small sample sizes (Arnau et al., 2013; Verbeke & Lesaffre, 1997). We did not apply robust estimation in our simulation and for the situations that we simulated here. However, we found that that ordinary estimation procedures perform sufficiently well. We feel encouraged by the observation that estimates are on average unbiased and are less concerned with the exact significance–non-significance cutoff that affects type I error rates. However, it is important to note that for individual cases, robust estimation might make a difference regarding whether the value is above or below some specific α threshold.

In our simulation, we only superficially investigated non-Gaussian error distributions, which can be modelled by generalized linear mixed-effects models (GLMMs). GLMMs include link functions and specific error distributions. The array of possible options is vast, with Poisson GLMMs, Binomial GLMMs and Negative-binomial GLMMs being the most popular. The link function and the specific error distributions cause additional complications and need to be chosen appropriately (Bolker et al., 2009; Harrison et al., 2018). However, at the latent scale, the models are equivalent to Gaussian GLMMs, such that most of our results will likely apply to non-Gaussian GLMMs as well. Our simulations for Poisson and proportion data indicate that results were broadly similar, but Poisson and Binomial models responded differently to violations in different situations. Robustness thus seems to be much more context-specific in the case of GLMMs. Furthermore, the expected response will also depend on the population mean value. For the simulation of binomial GLMMs we have chosen a mean of zero on the latent scale, which results in an expected

mean of 0.5 on the observed scale and thus maximum information per data point. Other cases may be less favourable.

Extensions of LMMs, such as double hierarchical generalized linear models (dHGLMs) allow modelling heterogeneous residual variances. This might be of specific biological interest (e.g. Westneat, Wright, & Dingemanse, 2015) and can be extended to fit fixed and random effect predictors that potentially affect heterogeneity in residual variances (rather than just affecting average trait values, Cleasby & Nakagawa, 2011; Cleasby, Nakagawa, & Schielzeth, 2015). Because of the small impact of heteroscedasticity on model estimates (see also Jacqmin-Gadda et al., 2007, but note the reduced coverage of the confidence interval for fixed effects), it does not seem to be necessary to fit heterogeneous residual variances when the main aim is to get robust estimates of fixed and random effects components in the fixed part of the model. dHGLMs make additional assumptions (including those about the distribution of heterogeneous residual variances) that will require additional checks. It seems that the biological insights promised by HGLMs, rather than putative violations per se, should be the main motivation for fitting them (Westneat et al., 2015).

In many cases, part of the appropriate model specification should include random-slope terms when data on slopes are replicated within groups (Gurka, Edwards, & Muller, 2011; Schielzeth & Forstmeier, 2009). Random-slope models can get complicated, in particular if multiple random-slope terms are to be included. In general, special attention is required such that models still report appropriate estimates. We did not simulate random-slope effects, in particular because the range of parameters to be explored goes beyond the scope of this manuscript. It is thus possible that violations of random effect distributions that are involved in random-slope terms might cause biased and/or imprecise estimates. Simulations of random-slope models with violated distribution of the random effects suggest robustness of the fixed effect type 1 error rate if the random-slope variance is appropriately specified (Jacqmin-Gadda et al., 2007; Taylor, Cumberland, & Sy, 1994; Warrington et al., 2014), but the conditions when random slopes are not appropriately specified are largely unexplored.

Overall, our results should be viewed as encouraging, and allow users of mixed-effects models to proceed with confidence. We conclude that mixed-effects models are largely robust even to quite severe violations of model assumptions. While it can be fine to model data that are clearly not normally distributed, we do caution that this might result in increased variability in estimates, hence extreme (and therefore misleading) results might occur for specific datasets. However, the effect is largely confined to the parameter(s) that are most closely linked to the violations of assumptions. It will thus be relatively easy to identify estimates that should be treated with increased caution. Our results therefore do not eliminate the need for proper evaluation of each model (of which posterior predictive model checks seem most generally applicable, Gelman & Hill, 2007). However, mixed-effects models should not be seen as dangerously complicated. Rather they are powerful tools to model a large variety of dataset and are usually more powerful than alternative analyses. We therefore encourage the use of mixed-effects models even for

slightly non-standard datasets while also advocating efforts to understand the specific consequences of particular types of violations.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest concerning the content of the manuscript.

AUTHORS' CONTRIBUTIONS

Idea: The analysis resulted from discussions within the SQuID (Statistical Quantification of Individual Differences) group. Simulation: H.S. data analysis: H.S. first draft of manuscript: H.S. All authors contributed to the discussion of results and to revisions in the manuscript. The order of co-authors was determined by a random number generator.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13434

DATA AVAILABILITY STATEMENT

Simulation scripts are available on Github https://github.com/hschielzeth/RobustnessGLMM (Schielzeth, 2020).

ORCID

Holger Schielzeth https://orcid.org/0000-0002-9124-2261
Niels J. Dingemanse https://orcid.org/0000-0003-3320-0861
Shinichi Nakagawa https://orcid.org/0000-0002-7765-5182
David F. Westneat https://orcid.org/0000-0001-5163-8096
Céline Teplitsky https://orcid.org/0000-0001-9458-709X
Denis Réale https://orcid.org/0000-0002-0419-7125
Ned A. Dochtermann https://orcid.org/0000-0002-8370-4614
László Zsolt Garamszegi https://orcid.org/0000-0001-8920-2183

REFERENCES

Allegue, H., Araya-Ajoy, Y. G., Dingemanse, N. J., Dochtermann, N. A., Garamszegi, L. Z., Nakagawa, S., … Westneat, D. F. (2017). Statistical quantification of individual differences (SQuID): An educational and statistical tool for understanding multilevel phenotypic data in linear mixed models. *Methods in Ecology and Evolution*, 8, 257–267. https://doi.org/10.1111/2041-210X.12659

Alonso, A., Litiere, S., & Laenen, A. (2010). A note on the indeterminacy of the random-effects distribution in hierarchical models. *The American Statistician*, 64, 318–324. https://doi.org/10.1198/tast.2010.09244

Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2013). The effect of skewness and kurtosis on the robustness of linear mixed models. *Behavior Research Methods*, 45, 873–879. https://doi.org/10.3758/s13428-012-0306-x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.

Becker, M., & Klößner, S. (2017). PearsonDS: Pearson distribution system. R package version 1.1. Retrieved from https://cran.r-project.org/web/packages/PearsonDS

Bolker, B. (2008). *Ecological models and data in R*. Princeton, NJ: Princeton University Press.

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. https://doi.org/10.1016/j.tree.2008.10.008

Box, G. E. P. (1980). Sampling and Bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society A*, 143, 383–430.

Brodie, E. D., Moore, A. J., & Janzen, F. J. (1995). Visualizing and quantifying natural selection. *Trends in Ecology & Evolution*, 10, 313–318. https://doi.org/10.1016/S0169-5347(00)89117-X

Cheng, J., Edwards, L. J., Maldonado-Molina, M. M., Komro, K. A., & Muller, K. E. (2010). Real longitudinal data analysis for real people: Building a good enough mixed model. *Statistics in Medicine*, 29, 504–520.

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. https://doi.org/10.1007/s00265-011-1254-7

Cleasby, I. R., Nakagawa, S., & Schielzeth, H. (2015). Quantifying the predictability of behaviour: Statistical approaches for the study of between-individual variation in the within-individual variance. *Methods in Ecology and Evolution*, 6, 27–37. https://doi.org/10.1111/2041-210X.12281

Demidenko, E., & Stukel, T. A. (2005). Influence analysis for linear mixed-effects models. *Statistics in Medicine*, 24, 893–909. https://doi.org/10.1002/sim.1974

Freckleton, R. P. (2011). Dealing with collinearity in behavioural and ecological data: Model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology*, 65, 91–101. https://doi.org/10.1007/s00265-010-1045-6

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

Grilli, L., & Rampichini, C. (2015). Specification of random effects in multilevel models: A review. *Quality & Quantity*, 49, 967–976. https://doi.org/10.1007/s11135-014-0060-5

Gurka, M. J., Edwards, L. J., & Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine*, 30, 2696–2707. https://doi.org/10.1002/sim.4293

Gurka, M. J., Edwards, L. J., Muller, K. E., & Kupper, L. L. (2006). Extending the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society A*, 169, 273–288. https://doi.org/10.1111/j.1467-985X.2005.00391.x

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., … Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, e4794. https://doi.org/10.7717/peerj.4794

Heagerty, P. J., & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, *88*, 973–985. https://doi.org/10.1093/biomet/88.4.973

Houle, D., Pelabon, C., Wagner, G. P., & Hansen, T. F. (2011). Measurement and meaning in biology. *The Quarterly Review of Biology*, *86*, 3–34. https://doi.org/10.1086/658408

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J. M., & Thiebaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics & Data Analysis*, *51*, 5142–5154. https://doi.org/10.1016/j.csda.2006.05.021

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997. https://doi.org/10.2307/2533558

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*, 2583–2595. https://doi.org/10.1016/j.csda.2008.12.013

Lee, Y., & Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, *19*, 219–228. https://doi.org/10.1214/088342304000000305

Loy, A., & Hofmann, H. (2013). Diagnostic tools for hierarchical linear models. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*, 48–61. https://doi.org/10.1002/wics.1238

Loy, A., & Hofmann, H. (2014). HLMdiag: A suite of diagnostics for hierarchical linear models in R. *Journal of Statistical Software*, *56*, 1–18.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137. https://doi.org/10.1046/j.0039-0402.2003.00252.x

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting It wrong may not matter. *Statistical Science*, *26*, 388–402. https://doi.org/10.1214/11-STS361

Morrissey, M. B., & Ruxton, G. D. (2018). Multiple regression is not multiple regressions: The meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory, and Practice in Biology*, *10*, 3. https://doi.org/10.3998/ptpbio.16039257.0010.003

Nakagawa, S. (2015). Missing data: Mechanisms, methods, and messages. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), *Ecological statistics: Contemporary theory and application* (pp. 81–105). Oxford, UK: Oxford University Press.

Nakagawa, S., & Freckleton, R. P. (2008). Missing inaction: The dangers of ignoring missing data. *Trends in Ecology & Evolution*, *23*, 592–596. https://doi.org/10.1016/j.tree.2008.06.014

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews*, *85*, 935–956. https://doi.org/10.1111/j.1469-185X.2010.00141.x

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge, UK: Cambridge University Press.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Royall, R. M. (1986). Model robust confidence-intervals using maximum-likelihood estimators. *International Statistical Review*, *54*, 221–226. https://doi.org/10.2307/1403146

Royall, R., & Tsou, T. S. (2003). Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions. *Journal of the Royal Statistical Society B*, *65*, 391–404. https://doi.org/10.1111/1467-9868.00392

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*, 1151–1172. https://doi.org/10.1214/aos/1176346785

Santos Nobre, J., & Singer, J. D. (2007). Residual analysis for linear mixed models. *Biometrical Journal*, *49*, 863–875. https://doi.org/10.1002/bimj.200610341

Schielzeth, H. (2020). Code for: Robustness of linear mixed-effects models to violations of distributional assumptions. *Zenodo*, https://doi.org/10.5281/zenodo.3877130

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*, 416–420. https://doi.org/10.1093/beheco/arn145

Sinharay, S., & Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, *111*, 209–221. https://doi.org/10.1016/S0378-3758(02)00303-8

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modelling* (2nd ed.). London, UK: Sage.

Taylor, J. M. G., Cumberland, W. G., & Sy, J. P. (1994). A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*, *89*, 727–736. https://doi.org/10.1080/01621459.1994.10476806

Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society Series C*, *37*, 87–94. https://doi.org/10.2307/2347496

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*, 217–221. https://doi.org/10.1080/01621459.1996.10476679

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, *23*, 541–556. https://doi.org/10.1016/S0167-9473(96)00047-3

Warrington, N. M., Tilling, K., Howe, L. D., Paternoster, L., Pennell, C. E., Wu, Y. Y., & Briollais, L. (2014). Robustness of the linear mixed effects model to error distribution assumptions and the consequences for genome-wide association studies. *Statistical Applications in Genetics and Molecular Biology*, *13*, 567–587. https://doi.org/10.1515/sagmb-2013-0066

Westneat, D. F., Wright, J., & Dingemanse, N. J. (2015). The biology hidden inside residual within-individual phenotypic variation. *Biological Reviews*, *90*, 729–743. https://doi.org/10.1111/brv.12131

Zare, K., & Rasekh, A. (2011). Diagnostic measures for linear mixed measurement error models. *Sort-Statistics and Operations Research Transactions*, *35*, 125–144.

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. https://doi.org/10.1111/j.2041-210X.2009.00001.x

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*. London, UK: Springer.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Schielzeth H, Dingemanse NJ, Nakagawa S, et al. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol Evol*. 2020;11:1141–1152. https://doi.org/10.1111/2041-210X.13434