

Including random effects in statistical models in ecology: fewer than five levels? (#60006)

1

First submission

Guidance from your Editor

Please submit by **7 May 2021** for the benefit of the authors (and your \$200 publishing discount) .



Structure and Criteria

Please read the 'Structure and Criteria' page for general guidance.



Raw data check

Review the raw data.



Image check

Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the [materials page](#).

5 Figure file(s)

1 Table file(s)




Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. BASIC REPORTING
2. EXPERIMENTAL DESIGN
3. VALIDITY OF THE FINDINGS
4. General comments
5. Confidential notes to the editor






 You can also annotate this PDF and upload it as part of your review

When ready [submit online](#).





Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your [guidance page](#).





BASIC REPORTING

-  Clear, unambiguous, professional English language used throughout.
-  Intro & background to show context. Literature well referenced & relevant.
-  Structure conforms to [Peerj standards](#), discipline norm, or improved for clarity.
-  Figures are relevant, high quality, well labelled & described.
-  Raw data supplied (see [Peerj policy](#)).

EXPERIMENTAL DESIGN

-  Original primary research within [Scope of the journal](#).
-  Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
-  Rigorous investigation performed to a high technical & ethical standard.
-  Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

-  Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
-  All underlying data have been provided; they are robust, statistically sound, & controlled.
-  Speculation is welcome, but should be identified as such.
-  Conclusions are well stated, linked to original research question & limited to supporting results.



The best reviewers use these techniques

Tip

Example

Support criticisms with evidence from the text or from other sources

Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.

Give specific suggestions on how to improve the manuscript

Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).

Comment on language and grammar issues

The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.

Organize by importance of the issues, and number your points

1. Your most important issue
2. The next most important item
3. ...
4. The least important points

Please provide constructive criticism, and avoid personal opinions

I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC

Comment on strengths (as well as weaknesses) of the manuscript

I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.

Including random effects in statistical models in ecology: fewer than five levels?

Dylan G.E. Gomes ^{Corresp. 1, 2}

¹ Biological Sciences, Boise State University, Boise, Idaho, United States

² Cooperative Institute for Marine Resources Studies, Hatfield Marine Science Center, Oregon State University, Newport, Oregon, United States

Corresponding Author: Dylan G.E. Gomes
Email address: dylangomes@u.boisestate.edu

As generalized linear mixed-effects models (GLMMs) have become a widespread tool in ecology, the need to guide the use of such tools is increasingly important. One common guideline is that one needs at least five levels of a random effect. Having such few levels makes the estimation of the variance of random effects terms (such as ecological sites, individuals, or populations) difficult, but it need not muddy one's ability to estimate fixed effects terms – which are often of primary interest in ecology. Here, I simulate ecological datasets and fit simple models and show that having too few random effects terms does not influence the parameter estimates or uncertainty around those estimates for fixed effects terms. Thus, it should be acceptable to use fewer levels of random effects if one is not interested in making inference ^S about the random effects terms (i.e. they are "nuisance" parameters used to group non-independent data). I also use simulations to assess the potential for pseudoreplication in (generalized) linear models (LMs) ² when random effects are explicitly ignored and find that LMs do not show increased type-I errors compared to their mixed-effects model counterparts. Instead, LM uncertainty (and p values) appears to be more conservative in an analysis with a real ecological dataset presented here. These results challenge the view that it is never appropriate to model random effects terms with fewer than five levels – specifically when inference is not being made for the random effects, but suggest that in simple cases LMs might be robust to ignored random effects terms. Given the widespread accessibility of GLMMs in ecology and evolution, future simulation studies and further assessments of these statistical methods are necessary to understand the consequences of both violating and ^{of} blindly following simple guidelines.

cf Hurlbert, Schielzeth + Forstmeier
"degrees of freedom" issues?

grouping variable?

Title:

Including random effects in statistical models in ecology: fewer than five levels?

Dylan G.E. Gomes^{1,2}

¹ Department of Biological Sciences, Boise State University, 1910 University Dr., Boise, ID, United States, 83725

² Cooperative Institute for Marine Resources Studies, Hatfield Marine Science Center, Oregon State University, 2030 SE Marine Science Dr., Newport, OR, United States, 97365

Email for correspondence: gomesd@oregonstate.edu

11 Abstract

12 As generalized linear mixed-effects models (GLMMs) have become a widespread tool in
 13 ecology, the need to guide the use of such tools is increasingly important. One common
 14 guideline is that one needs at least five levels of a random effect. Having such few levels makes
 15 the estimation of the variance of random effects terms (such as ecological sites, individuals, or
 16 populations) difficult, but it need not muddy one's ability to estimate fixed effects terms – which
 17 are often of primary interest in ecology. Here, I simulate ecological datasets and fit simple
 18 models and show that having too few random effects terms does not influence the parameter
 19 estimates or uncertainty around those estimates for fixed effects terms. Thus, it should be
 20 acceptable to use fewer levels of random effects if one is not interested in making inference
 21 about the random effects terms (i.e. they are 'nuisance' parameters used to group non-
 22 independent data). I also use simulations to assess the potential for pseudoreplication in
 23 (generalized) linear models (LMs), when random effects are explicitly ignored and find that LMs
 24 do not show increased type-I errors compared to their mixed-effects model counterparts. Instead,
 25 LM uncertainty (and p values) appears to be more conservative in an analysis with a real
 26 ecological dataset presented here. These results challenge the view that it is never appropriate to
 27 model random effects terms with fewer than five levels – specifically when inference is not
 28 being made for the random effects, but suggest that in simple cases LMs might be robust to
 29 ignored random effects terms. Given the widespread accessibility of GLMMs in ecology and
 30 evolution, future simulation studies and further assessments of these statistical methods are
 31 necessary to understand the consequences of both violating and blindly following simple
 32 guidelines.

33

34 **Keywords**

35 Statistics, hierarchical modelling, experimental design, block-design, varying effects,
 36 quantitative, regression, ANOVA

37 Introduction

38 As ecological datasets are inherently messy and researchers gain increased access to data,
 39 statistical analyses in ecology are becoming more complex (Low-Décarie, Chivers & Granados,
 40 2014), and advances in computing power and freely available statistical software are increasing
 41 the accessibility of such analyses to non-statisticians (Bates et al., 2007; Patil, Huard &
 42 Fonnesebeck, 2010; Gabry & Goodrich, 2016; Salvatier, Wiecki & Fonnesebeck, 2016; Bürkner,
 43 2017; Carpenter et al., 2017; Magnusson et al., 2017; Rue et al., 2017). As these methods have
 44 become more complex and accessible to ecologists, fisheries and wildlife managers, and
 45 evolutionary biologists the need to guide the use of such tools is becoming increasingly
 46 important (Bolker, 2008; Bolker et al., 2009; Zuur, Ieno & Elphick, 2010; Kéry & Royle, 2015;
 47 Kass et al., 2016; Zuur & Ieno, 2016; Harrison et al., 2018; Arnqvist, 2020; Silk, Harrison &
 48 Hodgson, 2020). ~~The use of~~ generalized linear mixed-effects models (GLMM), for example, ^{have}
 49 become a widespread tool that allows one to build hierarchical models that can estimate, and thus
 50 account for, imperfect detection in biological surveys (e.g. occupancy, N-mixture, mark-
 51 recapture, etc.) and can model correlations among data that come from non-independent groups
 52 or populations (i.e. random effects; also known as varying effects) (Bolker, 2008; Kéry & Royle,
 53 2015; Powell & Gale, 2015; Harrison et al., 2018; McElreath, 2020).

54

55 Generalized linear mixed-effects models are a regression type analysis that are flexible in that
 56 they can handle a variety of data generating processes such as binomial (e.g. presence / absence
 57 of a species, alive / dead individuals) and Poisson (e.g. wildlife counts). When the sampling
 58 distribution is Gaussian (also known as normal; e.g. mean-centered continuous data such as: tree

these are
extensions of
GLMMs

(conditional/
response) ~

terminology:
or 'response' (conditional)
distributions?

sampling distrib
of what?

diameter, vocalization frequency, or body condition residuals), this is a special case of a GLMM that is referred to as simply a linear mixed-effects model (LMM). GLMMs (and LMMs) differ from their simpler counterparts, (generalized) linear models (GLMs and LMs), in that they include random effects, in addition to the fixed effects (hence *mixed-effects*).

technically need real-valued response
OR positive vble with $CV \ll 1$

Fixed effects are also often called predictors, covariates, explanatory or independent variables in ecology and include both variables of interest (e.g. average temperature in climate change studies or sound pressure levels in anthropogenic noise research) and other variables that are only included to control for unexplained variation but ~~are~~ not directly useful to understanding the research question at hand (e.g. date or time of sampling in the above studies). Fixed effects are *fixed* in that the model parameters (β in equation 1 below) are assumed to be fixed, or non-random, and are not drawn from a hypothetical distribution.

why ital?

"nuisance"

are

Random effects, on the other hand, allow one to combine information (e.g. in a meta-analysis), deal with spatiotemporal autocorrelation, use partial pooling to borrow strength from other populations or groups, account for grouping or blocked designs (e.g. repeat-measures data from sites or individuals), and estimate population-level parameters, among others (Kéry & Royle, 2015). Thus, the random effects structure should be decided by the experimental design (Barr et al., 2013; Arnqvist, 2020). Random effects are *random* in that they are assumed to be drawn randomly from a distribution – often a Gaussian distribution – during the data-generating process. This is most often done by fitting a random *intercept* for each group (see equation 1 below), but one can, and should, also assign random *slopes* to variables, where the slopes of

applications

? their effects?

variables (not just the intercepts) are allowed to vary by group (see Bolker, 2008; Schielzeth & Forstmeier, 2009; Kéry & Royle, 2015; Harrison et al., 2018). If we are interested in the variability of a population (of individuals, groups, sites, or populations), it is difficult to estimate this variation with too few levels of individuals, groups, sites, or populations (i.e. random effects terms).

“When the number of groups is small (less than five, say), there is typically not enough information to accurately estimate group-level variation” (Gelman & Hill, 2006).

“...if interest lies in measuring the variation among random effects, a certain number is required...To obtain an adequate estimate of the among-population heterogeneity – that is, the variance parameter – at least 5 - 10 populations might be required” (Kéry & Royle, 2015).

“With <5 levels, the mixed model may not be able to estimate the among-population variance accurately” (Harrison et al., 2018).

“Strive to have a reasonable number of levels (at the very least, say, four to five subjects) of your random effects within each group” (Arnqvist, 2020).

This guideline that random effects terms should have at least five levels (i.e. groups) is backed by only limited empirical evidence (Harrison, 2015), but it is intuitive that too few draws from distribution will hinder one’s ability to estimate the variance of that distribution. Indeed, in each of the above segments of quoted text, the authors suggest that five levels are needed for *estimation of group-level, or among-population, variance*. However, this rule is often adhered to out of context, where authors or reviewers of ecological journals suggest that one cannot use random effects terms if they do not contain at least five levels, *in any case*. ✓

103

104 Simulations by Harrison (2015) demonstrate that random effects variance can be biased more
 105 strongly when the levels of random effects terms are low, yet in this work it appears that slope
 106 (beta) estimates for fixed effects terms are generally not more biased with only three random
 107 effects levels compared to five. There are many cases (and some would argue that in *most cases*,
 108 see below) in which the variance of random effects is not directly of interest to the ecological
 109 research question at hand.

estimates of

s

of direct ecological interest?

110 “...in the vast majority of examples of random-effects (or mixed) models in ecology, the
 111 random effects do *not* have a clear ecological interpretation. Rather, they are merely
 112 abstract constructs invoked to explain the fact that some measurements are more similar
 113 to each other than others are – i.e., to model correlations in the observed data” (Kéry &
 114 Royle, 2015).

115 Thus, it is unclear whether or not it is appropriate to use random effects when there are fewer
 116 than five grouping levels in situations where one does not directly care about the ‘nuisance’
 117 among-population variance, but instead is interested in estimates and uncertainty of predictor
 118 variables (i.e. fixed effects). The current state of practice in ecology is to drop the random effects
 119 terms such that we are now using generalized linear models where we are not grouping
 120 observations (we drop the **Mixed-effects** from the GLMM to become GLM). I question whether
 121 we are choosing to accept pseudoreplication in ecology (Hurlbert, 1984; Kéry & Royle, 2015;
 122 Arnqvist, 2020), rather than inaccurate estimates of among-population variance. In cases where
 123 one does not care about among-population variance, this tradeoff may be non-existent, but little
 124 research exists to support this ?

*or, treat
as fixed...*

125

126 Here, I simulate ecological datasets to assess whether *fixed effects* estimates are more biased
 127 when the accompanying *random effects* consist of fewer than five levels; I also ask whether
 128 using an alternative model without random effects (LMs) leads to higher type I errors
 129 (demonstrating a ‘significant’ effect when in fact one does not exist). I also analyze a real dataset
 130 within a similar framework to understand how the number of random effects levels and model
 131 structure (random effects or not) might change ecological inference.

132

133

focus on
coverage??

134 Methodology

135 All simulation of datasets and model fitting was done in R v4.0.4 (R Core Team, 2017). ^{all}
 136 visualizations were accomplished with the aid of R package ‘ggplot2’ (Wickham, 2011).

137

138 Data generation

139 I used a modified version of code from Harrison (2015), to explore the importance of varying
 140 two parameters in a linear mixed-effect model (LMM): the number of observations in a dataset
 141 (30, 60, or 120), and the number of levels of the random intercept term (3, 5, or 20). We can
 142 think of the latter as the number of individuals in an experiment or the number of field sites in a
 143 study. This was done by generating a response variable y_i from the following equation:

$$y_i = \alpha_{j(i)} + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \varepsilon_i$$

[1]

$$\alpha_j \sim \text{Normal}(\mu, \sigma)$$

[2]

Where $\alpha_{j(i)}$ is the intercept for site (or individual) j to which observation i belongs. Thus, each observation shared a site-level intercept, which ~~were~~ ^{was} drawn from a normal distribution with mean (μ) = 0 and standard deviation (σ) = 0.5. β_1 and β_2 are the slope parameters for two generic predictor variables (X_{1_i} and X_{2_i} respectively), which were both randomly generated from a normal distribution with $\mu = 0$ and $\sigma = 0.5$, which mimics standardized variables that are centered by their mean and scaled by two standard deviations (Gelman, 2008). Note that while X_{1_i} and X_{2_i} were drawn from a normal distribution during data generation, their associated parameters β_1 and β_2 are not (these are the *fixed* effects). For all simulated datasets, parameter values were fixed at $\beta_1 = 2$ and $\beta_2 = 0$, meaning X_{2_i} ~~does not have a linear relationship with,~~ or ^{is independent of} ~~is only randomly related to,~~ the response variable y_i . This allows for an assessment of type-I error rate, since any significant p values for this β_2 slope parameter are erroneous. The error term ε_i ~~is~~ ^{the} unique to each observation i , ~~that~~ is drawn from a normal distribution with $\mu_\varepsilon = 0$ and $\sigma_\varepsilon = 0.25$ (same as equation 2 above); this term is simply adding noise to our system during the data generating process, but is modelled implicitly as residual variance in many generic GLMM functions. [?]

? Informal?
one?

164 As an ecological example, you might think of the above equations as

165
$$\text{tree height}_i = \text{Site}_{j(i)} + 2 * \text{sunlight}_{1_i} + 0 * \text{ocean salinity}_{2_i} + \varepsilon_i$$

166
$$\text{Site}_j \sim \text{Normal}(\mu, \sigma)$$
 [2.1]
167
168 prefer • or juxtaposition
for math
expressions
(* for computer code) [2.1]

169 Data from each site or forest patch ($\text{Site}_1, \text{Site}_2, \dots, \text{Site}_n$) are grouped by their own site-level
170 intercept, which is randomly drawn from a normal distribution; some sites have shorter trees on
171 average and other sites have taller trees on average, but the entire forest (where all sites were
172 randomly selected from; equation 2.1) has a mean height of μ and a variance σ . We might expect
173 sunlight to have a positive relationship with tree height ($\beta_1 = 2$) and ocean salinity to reasonably
174 have no relationship with tree height ($\beta_2 = 0$). We could just leave the ocean salinity term out of
175 the equation entirely, but this explicitly demonstrates that the parameter coefficient that we
176 estimate with the model should $\neq 0$.
formulation shows
equal

177

178 *Model fitting simulations*

179 For each of the nine combinations of scenarios (30, 60, or 120 observations by 3, 5, or 20 sites), I
180 simulated 10,000 datasets. Each dataset was fit with a linear mixed-effect model (LMM) and a
181 linear model (LM). All model fitting was done with R functions 'lmer' (LMM) or 'lm' (LM) in
182 the package 'lme4' or in 'base' R, respectively (Bates et al., 2007; R Core Team, 2017), and p
183 values for 'lme4' models were calculated with 'lmerTest' (Kuznetsova, Brockhoff &
184 Christensen, 2017) (see note in discussion about this).

monospace font?

```

185 #LMM:
186 m1 <- lmer(y ~ x1 + x2 + (1|Site))
187 m1 <- lmer(tree.height ~ sunlight + ocean.salinity + (1|Site))

```

data =
(best practice)

188 R Code

189 Where x_1 and x_2 are fixed effects (see equation 1), and $(1|Site)$ is the syntax for specifying a
 190 random intercept ($\alpha_{j(i)}$ in equation 1). In ecology, we often fit independent sites as unique levels
 191 of a random effect, so I use site here for demonstration purposes. But site can be replaced with
 192 individual, group, population, etc. Often the recommendation, if one has fewer than 5 levels of
 193 random effects terms ($j < 5$ in $\alpha_{j(i)}$), is to fit the random effects as fixed effects (LMM becomes
 194 LM), specified in R as:

```

195 #LM:
196 m2 <- lm(y ~ x1 + x2 + Site)
197 m2 <- lm(tree.height ~ sunlight + ocean.salinity + Site)

```

198 R Code

199 and mathematically defined as:

$$y_i = \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 Site_{1(i)} + \beta_4 Site_{2(i)} + \dots \beta_{n+2} Site_{n(i)} + \varepsilon_i$$

201 [3]

202 Now a β term is estimated for each site (or population) level independently. Site effects no
 203 longer come from a normal distribution (as in equation 2), but instead are considered fixed,
 204 hence *fixed effects*. Thus, both a LMM and a LM were fit to each simulated dataset ($n = 10,000$)

of each of the nine combinations of data-generation (30, 60, or 120 observations by 3, 5, or 20 sites; 90,000 total simulated datasets and 180,000 models fit to data). This allowed for a comparison of the type-I error rates of LMMs and LMs, the latter of which ignores the blocked structure of data (i.e. site-level grouping).

Type-I error calculation and p values

Type-I error rate was calculated as the proportion of 10,000 models that a ‘significant’ p value of ≤ 0.05 was obtained for the β_2 parameter estimate (e.g. ocean salinity) in which the true value of that parameter was set to be 0. I sampled (with replacement) 10,000 p value ‘observations’ from each group of 10,000 models to produce a new proportion of type-I error; this process was repeated 1,000 times, and the bootstrapped 95% confidence intervals were calculated as the 0.025 and 0.975 quantiles of those 1,000 replications (see code; modified from Harrison, 2015).

shouldn't
binomial CIs
be OK?

Case study: spider body condition and noise

I used orb-weaving spider body condition data from a previous experiment (Gomes, Hesselberg & Barber, 2021). ~~In short,~~ speakers were used to experimentally broadcast whitewater river noise (predictor variable: sound pressure level) over the course of multiple summers. Orb-weaving spiders (*Tetragnatha versicolor*) were then weighed and measured (femur length), and body condition (response variable) was calculated from a residual index, using these measurements (Jakob, Marshall & Uetz, 1996; Gomes, 2020).

Here, I randomly sampled this dataset to create 30,000 new datasets, such that each new dataset contained three random observations from each of 3, 5, or 10 sites (10,000 datasets for each level of random effects terms). Thus, each dataset contained 9, 15, or 30 total observations when there were 3, 5, or 10 sites, respectively. While it would have been ideal to separate sample size from the number of random effects (as in the simulated datasets above), this simply wasn't possible to do with this real dataset (i.e. obtaining 30 total observations from three sites was not possible, while easily done for ten sites). Each dataset was fit with a linear mixed-effect model (LMM) and a linear model (LM) for comparison. Similar to the simulations above, all model fitting was done with R functions ``lmer`` (LMM) or ``lm`` (LM) in the package ``lme4`` or in ``base`` R, respectively (Bates et al., 2007; R Core Team, 2017) with the formulae:

```
m3 <- lmer(body.condition ~ sound.pressure.level + (1|Site)) # LMM
```

```
m4 <- lm(body.condition ~ sound.pressure.level + Site) # LM
```

bootstrap resds?

R Code

I used the same methodology as above in '*Type-I error calculation*' to calculate the proportion of 10,000 models in which $p < 0.05$, which corresponds to rejecting the null hypothesis that sound pressure level is not related to spider body condition. In this case, we are not assessing a type-I error, because we do not truly know if this variable should be significant or not, but instead it serves as a point of comparisons across methods (LMM vs LM) with real data rather than simulated.

Results

247 *Estimating model parameters and uncertainty*

re construct ?

248 Linear mixed models and linear models were able to resurrect simulated fixed effect
 249 relationships with no noticeable patterns in bias, regardless of number of levels of random effects
 250 or sample size. That is, both mean model parameter estimates (β_1 and β_2) were centered on their
 251 true values (2 and 0, respectively; Table 1; Figure 1, S1). The uncertainty around these estimates
 252 generally decreased as sample size increased. For example, doubling the sample size from 30
 253 observations to 60 observations lead to a decrease by 36.6% and 35.5% in parameter estimate
 254 uncertainty (for β_1 and β_2 respectively; Table 1; Figure 1). Another doubling to 120
 255 observations lead to a further decrease in uncertainty by 33.4% and 32.9%, respectively. The
 256 number of levels of random effects appears to be relatively non-important in resurrecting model
 257 parameter estimates within these simulation scenarios (Table 1; Figure 1); instead there were
 258 small, likely negligible, increases in uncertainty around fixed effect parameter estimates as the
 259 number of levels of random effects increased.

260

261 All LMM estimates of the distribution mean (μ) were unbiased, regardless of number of levels of
 262 random effects or sample size (Table 1; Figure 2A). The random effects variance (σ) estimates,
 263 however, were not centered at the true value, and estimates were more biased with fewer levels
 264 of random effects, whereas sample size did not affect this bias (Table 1; Figure 2B). That is, with
 265 only three levels of random effects the magnitude of the bias was 12.2% of the true value.
 266 Increasing to five levels of random effects nearly halved this bias to 6.4%, and increasing to 10
 267 levels halved the bias again to 3.2% of the true value. Averaged across numbers of random

effects terms, estimates were biased by about 7% regardless of sample size (7.1%, 7.4%, and 7.2% for $N = 30, 60$, and 120 respectively).

The uncertainty around random effects estimates (μ and σ) generally decreased with an increased number of random effects levels, whereas sample size did little to alleviate this uncertainty (Table 1; Figure 2). Increasing the number of random effects levels from 3 to 5, and then from 5 to 10, decreased the uncertainty for μ by 22.4% and 29.1%, respectively, and for σ by 26.6% and 29.8% respectively.

Type-I errors

For all simulated datasets, both LMM and LM produced type-I error rates around the typical $\alpha = 0.05$, with 95% confidence intervals overlapping this value. Neither sample size, nor the number of random effects levels seemed to influence the type-I error rate. Furthermore, dropping the random effects structure (using a LM instead of a LMM) did not increase the probability type-I errors (Figure 3), nor the ability of the model to accurately estimate fixed effects parameters (fixed effects estimates appear the same as they do in Figure 1, see Figure S1).

Case study results

Linear mixed models (LMM) and linear models (LM) both consistently estimated the fixed effect parameters for sound pressure level to be very weakly positive, with large overlap with zero (or no relationship). Linear model estimates are slightly shrunk towards zero, compared to

LMMs (Figure 4A), but it is impossible to assess bias, since we do not know truth with these biological data. The uncertainty around these estimates decreases with LMMs, compared to LMs, and decreases as the sample size and number of sites increases (in tandem).

Output from 10,000 LMs suggests that the null hypothesis (no effect) would be rejected by a proportion of approximately 0.05 (Figure 4B), which is around the typical type-I error alpha level (i.e. suggesting that sound pressure level is not a significant predictor of spider body condition). LMMs, on the other hand, experienced null hypothesis rejection rates at nearly double that of their LM counterparts, which is consistent with the reduced uncertainty around the LMM estimates in Figure 4A. The rate of rejecting the null hypothesis, however, did not appear to be related to the number of sites (and thus total observations).

Discussion

The work presented here demonstrates that i) fixed effects estimates are not more biased when the levels of an accompanying random effect have fewer than five ($n < 5$) levels, but population-level (random effects) variance estimates are and ii) type-I error rates are not increased by using linear models (LM) instead of linear mixed-effects models (LMM).

Fixed effects parameter estimation does not appear to be strongly influenced by, nor biased by, the number of levels of random effects terms. Instead, uncertainty in those estimates is much more strongly influenced by sample size. While this pattern may appear to contradict the

decreased uncertainty (with more random effects levels) around beta estimates in Figure 2 of Harrison (2015) and in Figure 4 of the current work, this instead is due to differences in the way that sample size relates to the number of random effects levels. Harrison (2015) coded each random effect level to be associated with a fixed number of observations ($N=20$), such that each additional random effect level yielded an increased sample size – as is also the case in the spider case study presented here. However, in the simulations here (Figure 1), sample size (i.e. number of observations) has been separated from the number of random effects terms (e.g. sites or individuals).

Despite these differences in coding, the estimation of random effects terms (μ and σ) in the simulations here suggest consistent patterns with Harrison (2015) in that variance (σ) is more biased with fewer levels. This seems to support previous suggestions and simulations suggesting that few levels of random effects terms can make estimation of population-level variance difficult (Gelman & Hill, 2006; Harrison, 2015; Kéry & Royle, 2015; Harrison et al., 2018), but the cutoff at five random effects levels appears quite arbitrary. The combination of these results suggest that using fewer than five levels of random effects is acceptable when one is only interested in estimating fixed effects parameters (i.e. predictors, independent variables); in other words, when inference about the variance of random effects terms (e.g. sites, individuals, populations) is not of direct interest, but instead are used to group non-independent data. In these cases, however, caution should be taken in reporting the variance estimates for such population-level parameters – as this information can later be taken out of context of the question at hand and may result in the propagation of biased estimates.

332

333 Those following the “less than five levels” guideline typically drop random effects from
 334 analyses, turning a LMM into a LM. In both the simulations and the case study, LMMs and LMs
 335 did not appear to give drastically different parameter estimates for fixed effects. In the spider
 336 case study, LMMs gave more consistent results, leading to increased parameter certainty when
 337 compared to LMs, which was also reflected in a higher probability of ‘significant’ p values
 338 (around 10% of the models), when compared to LMs (around 5% of the models, which is to be
 339 expected to due chance). That is, results from LMs here suggest that there is no significant effect
 340 of the predictor (sound pressure level) on the response (body condition). While misspecified
 341 mixed-effects models can be overconfident in their estimates (Schielzeth & Forstmeier, 2009), in
 342 this case we do not know what ‘truth’ is. That is, we do not know if we *should* be rejecting the
 343 null hypothesis since the biological data here are real. However, this highlights that the p values
 344 might not always be very informative. For both model types in the case study (LMM vs LM), the
 345 magnitude of the parameter estimates (i.e. effect sizes) were consistently small. Interpreting the
 346 estimates directly will likely lead to a more consistent understanding of the results, rather than
 347 focusing on whether p values pass an arbitrary threshold.

348

349 The use, and abuse, of p values, in general, is highly debated and controversial (Yoccoz, 1991;
 350 Schervish, 1996; Wagenmakers, 2007; Murdoch, Tsai & Adcock, 2008; Gelman, 2013;
 351 Murtaugh, 2014; Leek & Peng, 2015; Greenland et al., 2016; Ho et al., 2019), but is further
 352 complicated by mixed-effects models (Luke, 2017). Douglas Bates, and other authors of the R
 353 package ‘lme4’ (Bates et al., 2007) which I use here, does not include a p value ‘baked in’ to the

output for a reason (in short how one should calculate this is not straightforward or as obvious as it sounds). While my personal approach is generally to rely more on probabilistic (i.e. Bayesian) approaches or effect sizes and ignore p values wherever possible, I use p values here because they are still widely used across ecology and evolutionary studies and by fish and wildlife managers. The R package `'lmerTest'` (Kuznetsova, Brockhoff & Christensen, 2017) allows many users to get around this *apparent* limitation of `'lme4'` by calculating p values as if they were a part of the original package. Despite my recommendation to focus more on effect size and other model outputs in analyses of real data, understanding the consequences of type-1 errors in terms of p values is relevant so long as ecologists continue to use them.

Interestingly, type-I errors were not more likely in any of the LMs of simulated data. This possibly suggests that misspecified linear models (theoretically missing a random effect) are relatively robust to this omission – at least in some simple cases such as the scenarios presented here. While this perhaps alleviates some concern over inflated type-I errors due to pseudoreplication while ignoring the grouped nature of repeat-measures studies and non-independent data (Arnqvist, 2020), this should not be taken as evidence to purposefully omit random effects when such a structure is appropriate. Instead, it warrants future investigation and further simulation studies with more thorough scenarios (especially with varying degrees of random effect variance) and more complex data structures (e.g. including correlations and link functions).

Often researchers (sometimes nudged by peer-reviewers) cite this guideline of needing 5 levels before random effects inclusion as a reason why they were unable to use a mixed-effects model (Bain, Johnson & Jones, 2019; Bussmann & Burkhardt-Holm, 2020; Evans & Gawlik, 2020; Gomes & Goerlitz, 2020; Zhao, Johnson-Bice & Roth, 2021). Although there is confusion over this recommendation, as some opt to use mixed-effects models despite this suggestion (Latta et al., 2018; Fugère, Lostchuck & Chapman, 2020; Gomes, Appel & Barber, 2020; Allen et al., 2021), likely because of the numerous advantages that mixed-effects models offer (Bolker, 2008; Kéry & Royle, 2015; Harrison et al., 2018), or fear of the consequences of pseudoreplication (although this can easily occur in mixed-effects models as well: Schielzeth & Forstmeier, 2009; Arnqvist, 2020). The trend to automatically follow this rule is likely exacerbated by the fact that authors or peer-reviewers can easily point out that this rule exists (Gelman & Hill, 2006; Harrison, 2015; Kéry & Royle, 2015; Harrison et al., 2018; Arnqvist, 2020), but may find it more difficult or time-consuming to make a nuanced argument against following such a rapidly growing rule. Hopefully the results presented here will challenge that view, and allow the fitting of random effects when inference is not being made for the random effects. More importantly, I hope it sparks further conversation and debate over this issue. Given the widespread accessibility of GLMMs, future simulation studies and further assessments of these statistical methods are necessary to understand the consequences of both violating and blindly following methodological rules.

References

- Allen LC, Hristov NI, Rubin JJ, Lightsey JT, Barber JR. 2021. Noise distracts foraging bats. *Proceedings of the Royal Society B* 288:20202689.
- Arnqvist G. 2020. Mixed models offer no freedom from degrees of freedom. *Trends in ecology & evolution* 35:329–335.
- Bain GC, Johnson CN, Jones ME. 2019. Chronic stress in superb fairy-wrens occupying remnant woodlands: Are noisy miners to blame? *Austral Ecology* 44:1139–1149.
- Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68:255–278.
- Bates D, Sarkar D, Bates MD, Matrix L. 2007. The lme4 package. *R package version 2.*
- Bolker BM. 2008. *Ecological models and data in R*. Princeton University Press.
- Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* 24:127–135.
- Bürkner P-C. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80:1–28.
- Bussmann K, Burkhardt-Holm P. 2020. Round gobies in the third dimension-use of vertical walls as habitat enables vector contact in a bottom-dwelling invasive fish. *Aquatic Invasions* 15:683–699.
- Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Riddell A. 2017. Stan: a probabilistic programming language. *Grantee Submission* 76:1–32.
- Evans BA, Gawlik DE. 2020. Urban food subsidies reduce natural food limitations and reproductive costs for a wetland bird. *Scientific reports* 10:1–12.

- 420 Fugère V, Lostchuck E, Chapman LJ. 2020. Litter decomposition in Afrotropical streams:
421 Effects of land use, home-field advantage, and terrestrial herbivory. *Freshwater Science*
422 39:497–507.
- 423 Gabry J, Goodrich B. 2016. *rstanarm: Bayesian applied regression modeling via Stan. R*
424 *package version 2.10. 0*.
- 425 Gelman A. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in*
426 *medicine* 27:2865–2873.
- 427 Gelman A. 2013. Commentary: P values and statistical practice. *Epidemiology* 24:69–72.
- 428 Gelman A, Hill J. 2006. *Data analysis using regression and multilevel/hierarchical models*.
429 Cambridge University Press.
- 430 Gomes DGE. 2020. Orb-weaving spiders are fewer but larger and catch more prey in lit bridge
431 panels from a natural artificial light experiment. *PeerJ* 8:e8808.
- 432 Gomes DG, Appel G, Barber JR. 2020. Time of night and moonlight structure vertical space use
433 by insectivorous bats in a Neotropical rainforest: an acoustic monitoring study. *PeerJ*
434 8:e10591.
- 435 Gomes DGE, Goerlitz HR. 2020. Individual differences show that only some bats can cope with
436 noise-induced masking and distraction. *PeerJ* 8:e10551.
- 437 Gomes DGE, Hesselberg T, Barber JR. 2021. Phantom river noise alters orb-weaving spider
438 abundance, web size, and prey capture. *Functional Ecology* 35:717–726.
- 439 Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. 2016.
440 Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.
441 *European Journal of Epidemiology* 31:337–350.

- Harrison XA. 2015. A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution. *PeerJ* 3:e1114.
- Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CE, Robinson BS, Hodgson DJ, Inger R. 2018. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* 6:e4794.
- Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. 2019. Moving beyond P values: data analysis with estimation graphics. *Nature methods* 16:565–566.
- Hurlbert SH. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54:187–211.
- Jakob EM, Marshall SD, Uetz GW. 1996. Estimating fitness: a comparison of body condition indices. *Oikos*:61–67.
- Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N. 2016. *Ten simple rules for effective statistical practice*. Public Library of Science.
- Kéry M, Royle JA. 2015. *Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models*. Academic Press.
- Kuznetsova A, Brockhoff PB, Christensen RH. 2017. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software* 82:1–26.
- Latta SC, Brouwer NL, Mejía DA, Paulino MM. 2018. Avian community characteristics and demographics reveal how conservation value of regenerating tropical dry forest changes with forest age. *PeerJ* 6:e5217.
- Leek JT, Peng RD. 2015. Statistics: P values are just the tip of the iceberg. *Nature News* 520:612.

Low-Décarie E, Chivers C, Granados M. 2014. Rising complexity and falling explanatory power in ecology. *Frontiers in Ecology and the Environment* 12:412–418.

Luke SG. 2017. Evaluating significance in linear mixed-effects models in R. *Behavior research methods* 49:1494–1502.

Magnusson A, Skaug H, Nielsen A, Berg C, Kristensen K, Maechler M, van Bentham K, Bolker B, Brooks M, Brooks MM. 2017. Package ‘glmmTMB.’ *R Package Version 0.2. 0*.

McElreath R. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Murdoch DJ, Tsai Y-L, Adcock J. 2008. P-values are random variables. *The American Statistician* 62:242–245.

Murtaugh PA. 2014. In defense of P values. *Ecology* 95:611–617.

Patil A, Huard D, Fonnesbeck CJ. 2010. PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software* 35:1.

Powell LA, Gale GA. 2015. *Estimation of Parameters for Animal Populations*. Caught Napping Publications, Lincoln, NE.

R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK. 2017. Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application* 4:395–421.

Salvatier J, Wiecki TV, Fonnesbeck C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* 2:e55.

Schervish MJ. 1996. P values: what they are and what they are not. *The American Statistician* 50:203–206.

- Schielezeth H, Forstmeier W. 2009. Conclusions beyond support: overconfident estimates in mixed models. *Behavioral ecology* 20:416–420.
- Silk MJ, Harrison XA, Hodgson DJ. 2020. Perils and pitfalls of mixed-effects regression models in biology. *PeerJ* 8:e9522.
- Wagenmakers E-J. 2007. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review* 14:779–804.
- Wickham H. 2011. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* 3:180–185.
- Yoccoz NG. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* 72:106–111.
- Zhao S-T, Johnson-Bice SM, Roth JD. 2021. Foxes facilitate other wildlife through ecosystem engineering activities on the Arctic tundra. *bioRxiv*.
- Zuur AF, Ieno EN. 2016. A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution* 7:636–645.
- Zuur AF, Ieno EN, Elphick CS. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution* 1:3–14.

Table 1(on next page)

Model estimates from 10,000 simulated datasets.

The number of levels of random effects (RE) was varied (3, 5, or 10), as was the number of observations in the dataset ($N = 30, 60, \text{ or } 120$). The true (T) values for the data generation process (equation 1) are indicated in the second header row underneath the estimated parameter labels (fixed effects: β_1, β_2 ; random effects: μ, σ). The mean of 10,000 model estimates ($\beta_1, \beta_2, \mu, \sigma$) are indicated for the respective models below the true values. Lower and upper bounds on 95% confidence intervals for each parameter is calculated as the 0.025 and 0.975 quantiles, respectively, of 1,000 bootstrapped replications (see methods).

RE levels	N	β_1	β_1 95% CI		β_2	β_2 95% CI		μ	μ 95% CI		σ	σ 95% CI	
		T = 2	Lower	Upper	T = 0	Lower	Upper	T = 0	Lower	Upper	T = 0.5	Lower	Upper
3	30	1.999	1.792	2.202	0.001	-0.199	0.205	0.000	-0.580	0.589	0.443	0.000	0.982
5	30	2.001	1.790	2.217	0.001	-0.207	0.211	0.001	-0.450	0.454	0.468	0.147	0.846
10	30	2.002	1.764	2.241	-0.003	-0.237	0.230	-0.001	-0.322	0.324	0.482	0.245	0.739
3	60	2.001	1.864	2.139	0.000	-0.135	0.134	0.002	-0.573	0.569	0.437	0.053	0.964
5	60	2.000	1.865	2.136	0.000	-0.140	0.135	0.001	-0.443	0.441	0.466	0.161	0.831
10	60	2.001	1.858	2.143	0.000	-0.144	0.142	-0.002	-0.314	0.305	0.485	0.265	0.736
3	120	2.000	1.910	2.090	0.000	-0.093	0.091	-0.001	-0.565	0.568	0.438	0.067	0.954
5	120	2.000	1.907	2.093	0.000	-0.091	0.093	-0.001	-0.442	0.443	0.469	0.164	0.838
10	120	2.000	1.905	2.093	0.000	-0.094	0.096	0.000	-0.318	0.311	0.485	0.267	0.735

Figure 1

Fixed effects model estimates for simulated data.

Each point is the mean estimate for 10,000 models (and datasets), whereas error bars are 95% confidence intervals. N = the number of observations (i.e. number of rows) in each dataset. Dashed lines indicate the true value. In all scenarios the bias in parameter estimates are negligible. As the sample size increases, our certainty around the parameter estimates (β) increases, but the number of random effects has a relatively minor effect on estimating β .

When sample sizes (N) are low, parameter uncertainty increases with increasing levels of random effects (assuming a consistent N).

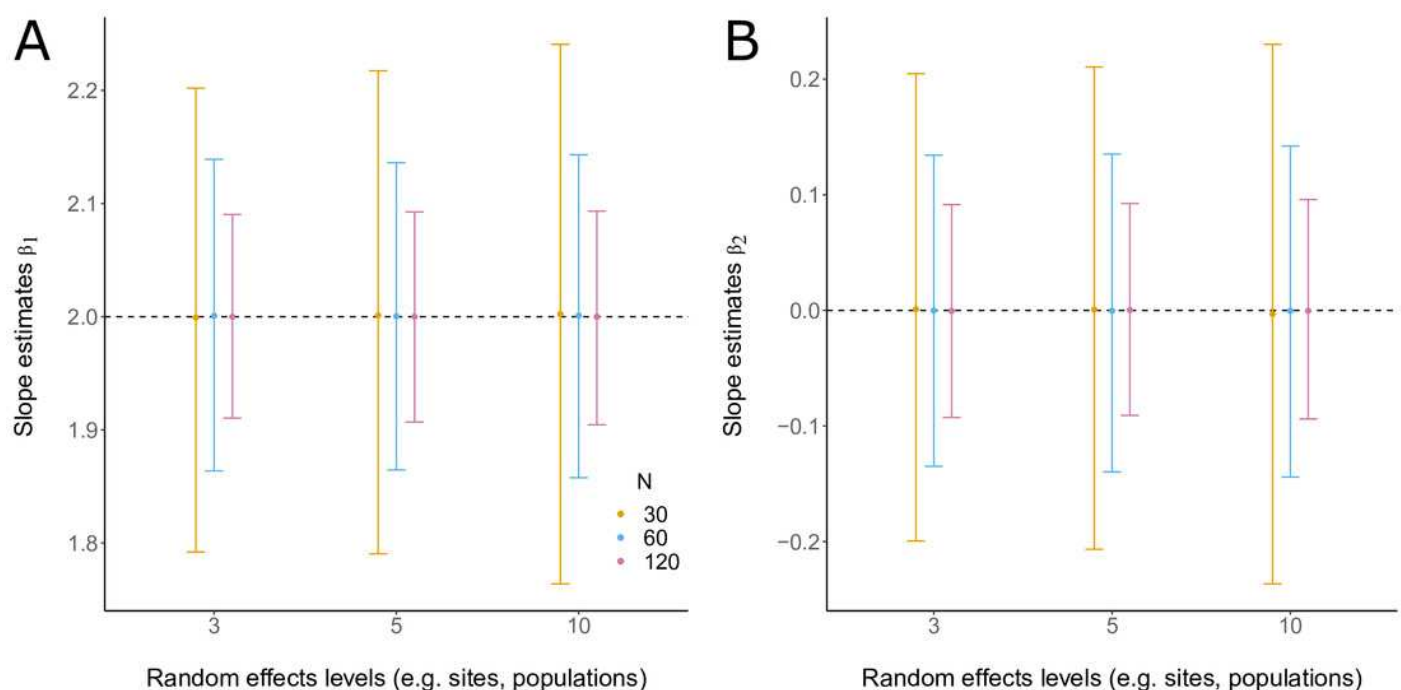


Figure 2

Random effects model estimates for simulated data.

Each point is the mean estimate for 10,000 models (and datasets), whereas error bars are 95% confidence intervals. N = the number of observations (i.e. number of rows) in each dataset. Dashed lines indicate the true value. A) As the number of random effects levels increases, the uncertainty around the mean (μ) decreases. Sample size has a relatively minor effect on estimating μ . B) As the number of random effects levels increases, the bias and uncertainty around the random effects variance (σ) decreases. Sample size has a small, but relatively minor effect on estimating σ . The bias in σ starts to approach the starting (simulated) $\sigma = 0.5$ as the number of random effects reaches 10.

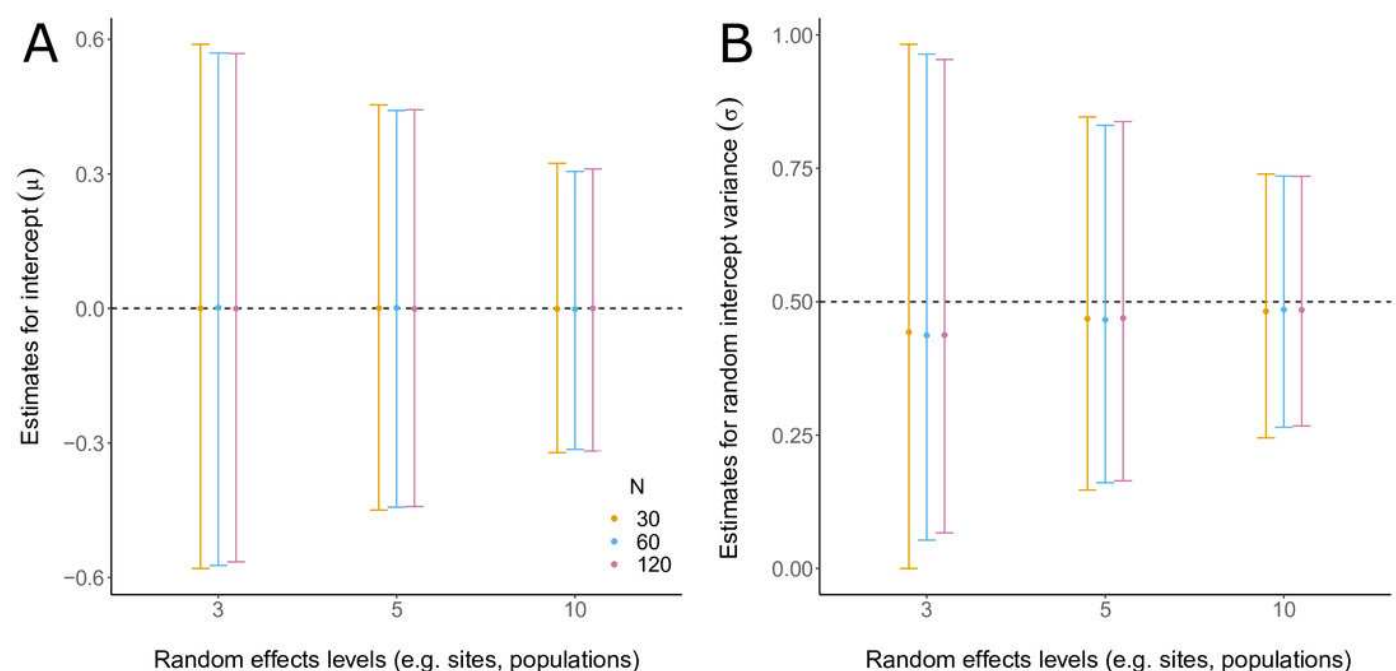


Figure 3

Type-I error for various linear models (LM) and linear mixed-effects models (LMM).

Type-I error rate was calculated as the proportion of models ($n = 10,000$) in which a 'significant' p value of ≤ 0.05 was obtained for a parameter estimate in which the true value of that parameter was set to be 0 (Figure 1B); each point represents this proportion. To generate error bars as 95% confidence intervals, I used bootstrapping to replicate this process 1,000 times (see methods). N = the number of observations (i.e. number of rows) in each dataset. Symbols indicate model type (LM vs LMM). Dashed lines indicate the true alpha value (0.05).

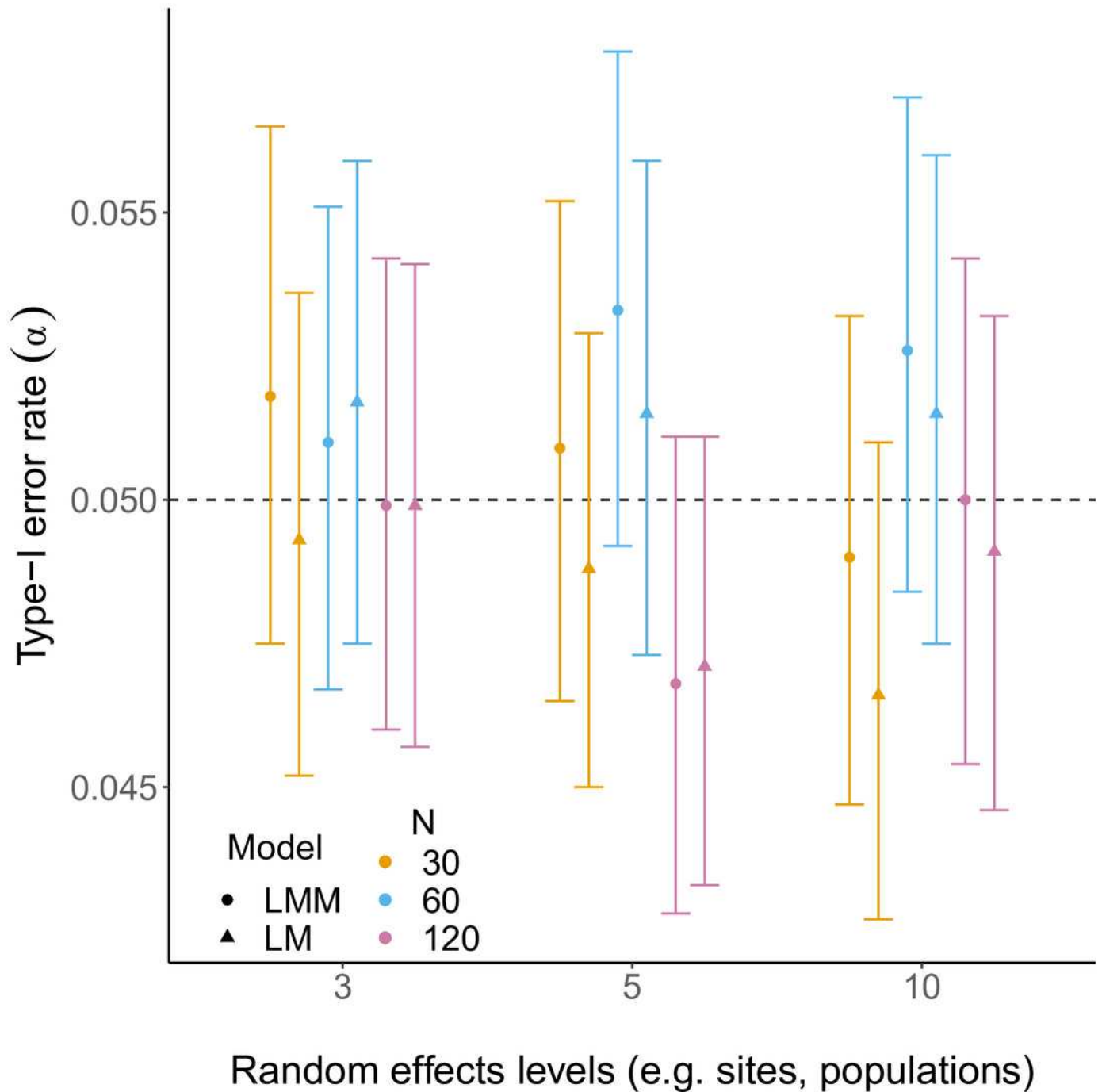


Figure 4

Analyses for real dataset (spiders in noise case study).

(A): Fixed effects model estimates do not differ strongly across number of sites or model types, but estimates from linear models are slightly pulled toward zero. Each point in is the mean estimate for 10,000 models (and datasets), whereas error bars are 95% confidence intervals. N = the number of observations (i.e. number of rows) in each dataset. Dashed line at zero indicates the hypothesis that there is no effect of sound pressure level on spider body condition. As the sample size and the number of sites increases, our certainty around the parameter estimates increases. This is likely due to sample size rather than the number of sites (see Figure 3). Estimates from linear models (LM) are slightly more uncertain than estimates from linear mixed-effects models (LMM). (B): The probability for rejecting the null ($p < 0.05$) is higher for LMMs than LMs, suggesting that LMs are more conservative in this scenario. The probability of rejecting the null model for LMs overlaps $\alpha = 0.05$, suggesting there is no effect of sound pressure level on spider body condition. Of course we do not know if rejecting the null model is “truth” here, since these are real biological data (see Figure 3 for simulated data). The probability of rejecting the null model was calculated as the proportion of models ($n = 10,000$) in which a ‘significant’ p value of ≤ 0.05 was obtained for the fixed effect parameter estimate; each point represents this proportion. To generate error bars as 95% confidence intervals, I used bootstrapping to replicate this process 1,000 times (see methods).

