Communications in Kinesiology

Methodological Tutorial

DOI:
10.51224/cik.2023.52

Subject Areas:
Metascience

Keywords:
mixed-effects regression, random effects, longitudinal data, repeated measures

Editor:
Matthieu Boisgontier

Model Specification in Mixed-Effects Models

A Focus on Random Effects

Keith R. Lohse[1,2,*], Allan J. Kozlowski[3], and Michael J. Strube[4]

[1]Program in Physical Therapy, Washington University School of Medicine
[2]Department of Neurology, Washington University School of Medicine
[3]Kozlowski Health Services Consulting LLC
[4]Department of Psychological and Brain Sciences, Washington University in Saint Louis
[*]Correspondence: lohse@wustl.edu

Mixed-effect models are flexible tools for researchers in a myriad of fields, but that flexibility comes at the cost of complexity and if users are not careful in how their model is specified, they could be making faulty inferences from their data. We argue that there is significant confusion around appropriate random effects to be included in a model given the study design, with researchers generally being better at specifying the fixed effects of a model, which map onto to their research hypotheses. To that end, we present an instructive framework for evaluating the random effects of a model in three different situations: (1) longitudinal designs; (2) factorial repeated measures; and (3) when dealing with multiple sources of variance. We provide worked examples with open-access code and data in an online repository. We think this framework will be helpful for students and researchers who are new to mixed effect models, and to reviewers who may have to evaluate a novel model as part of their review.

Introduction

Mixed-effect models (Fisher, 1919; McLean et al., 1991; Robinson, 1991) are an increasingly popular analytical method in a wide range of disciplines. As a crude illustration of its popularity, we conducted a literature search in PubMed to identify articles that reported using an Analysis of Variance (ANOVA) versus mixed-effect models. While ANOVA was by the far the dominantly used term, various references to mixed-effect models (sometimes also called mixed-effects regression, mixed models, or multi-level models (Raudenbush & Bryk, 2002; Singer & Willett, 2003)) have been closing the gap. This increase in the use of mixed-effect models is multi-causal, but broadly we think that researchers are attracted to mixed-effect models for their flexibility: mixed-effect models can deal with missing data more easily; mixed-effect models highlight individual differences in the estimation of variance components; for longitudinal designs, mixed-effect models explicitly model individual trajectories and variability in time itself; and mixed-effect models allow for more than one source of variability to be estimated simultaneously (Garcia & Marder, 2017; Lohse et al., 2020; Long, 2012; Snijders & Bosker, 2012). The mixed-effect model also extends to other statistical applications such as meta-analysis (Sera et al., 2019), measurement (Van der Elst et al., 2016), and mediation models (Kenny et al., 2003; Preacher et al., 2010). Finally, as with general linear models, mixed-effect models can be extended to handle categorical outcome measures (Agresti, 2015).

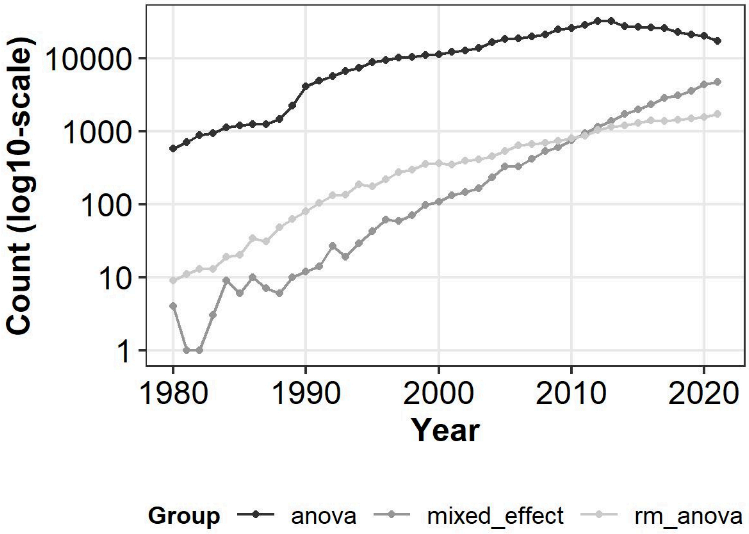


Figure 1. Frequency of hits on PubMed searched 2021-11-15 for “ANOVA”, “mixed-effect/s” (or “mixed effect/s”) and “repeated measures+ANOVA” (or “RM” + “ANOVA”).

The tremendous power of mixed models comes at the cost of analytical complexity. Despite excellent resources on the topic (Long, 2012; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012), researchers are often faced with difficult choices and sometimes conflicting recommendations (Barr, 2013; Bates, Kliegl, et al., 2015), and may not have the time or statistical expertise necessary to understand these resources (Silk et al., 2020), as many references are quite mathematically dense. Although there are domain-specific guides that are written to address particular types of research designs and use examples familiar to subject-matter experts in that domain (e.g., B. M. Bolker et al. (2009) and Brown (2021)), confusion surrounds this complicated topic.

Our own personal experience, which includes a non-random sample of dissertation committees, manuscript reviews, and questions at workshops on mixed-effect models, shows that a major part of this confusion is in choosing the appropriate random effects for a model, with models often being mis-specified. Researchers generally appear better at selecting and interpreting appropriate fixed effects to test their hypotheses. These fixed effects correspond most often to the hypotheses that guided the research. However, the inferences about fixed effects depend on the random effects of the model.

Inappropriate specification of random effects has substantial negative consequences on the credibility of the statistical inference (Frossard & Renaud, 2019). For instance, including only a random effect of participant in a study with multiple within-subject factors can make it seem like the researchers have more independent data than they have, spuriously shrinking the standard error and increasing the Type I error rate (an example of under-specification). Conversely, including fully crossed random effects of participants and stimuli in an experiment where there is only one observation per person for each stimulus can leave no residual error, creating a non-identifiable matrix (an example of over-specification). Or the researchers may simply omit a random-effect that is justified either by the design or the empirical variability in the data (an example of mis-specification). Thus, reasonable specification of random effects is critical to making reasonable inferences about the fixed effects (Faraway, 2016; Pinheiro & Bates, 2006; Singmann & Kellen, 2019).

Our primary contention is that there is confusion among researchers about how to correctly specify mixed-effect models (as Frossard & Renaud (2019), Nelder (2007), Voss (1999), and Senn (2003) have argued). Our goal is thus to provide a technically sound but accessible and didactic paper to help researchers choose appropriate random effects as a starting point based on their study design. Our target audience is busy researchers who are not experts: such as, novices who are starting out with mixed-effect models, researchers who are familiar with the basic concept but unsure how to properly specify a model in a more complicated design, and reviewers tasked with evaluating a novel model during peer-review.

We will not be able to move someone from novice to expert with the concepts presented in this manuscript, but we can help researchers avoid some fundamental errors and prepare them to tackle more detailed texts (Bates, Machler, et al., 2015; Faraway, 2016; Pinheiro & Bates, 2006). Our objective is to get researchers to think more broadly and carefully about random effects and the need to treat them appropriately in model specification. We also assume that all readers will be familiar with the language of “traditional” ordinary least squares regression (e.g., normality of residuals, contrast versus treatment coding, types of sums of squared errors, collinearity). If any of these terms are unfamiliar, we recommend studying those issues first (Cohen et al., 2003; Fox, 2016; Fox & Weisburg, 2011). Mixed-effect models are an extension of the general linear model and as such starting with simpler models is very helpful for tackling more complex designs.

Finally, we want to address some issues we will not cover. Our focus is on selecting a reasonable random-effects structure given the study design. As we will see this is more of a starting point and there is no universally correct structure (Barr, 2013; Bates, Kliegl, et al., 2015; Frossard & Renaud, 2019), but there are some common incorrect structures that authors should avoid. Our manuscript is therefore not a general introduction to mixed-effect models (Brown (2021); Singmann & Kellen (2019)), nor we will discuss the intricacies of how one should specify fixed effects, interpret parameters in the presence of interactions, or select between competing models (Faraway, 2016; Long, 2012; Pinheiro & Bates, 2006; Snijders & Bosker, 2012). We will also not address advanced concepts like statistical power in mixed-effect regression (Westfall et al., 2014) nor the application of Bayesian estimation to mixed-effect regression (Bates, Kliegl, et al., 2015). With that in mind, we first define notation and terms then address common designs as follows: 1. Longitudinal Designs, 2. Repeated Measures Factorial Designs, and 3. Designs with Multiple Sources of Variance. We try to keep our discussion of these issues concise, but we provide vignettes with worked examples in an evolving online repository.

Notation and Terms

In the manuscript, we focus on using R code (2023) and specifically `lme4` (Bates, Machler, et al., 2015) and `nlme` (Pinheiro et al., 2022) syntax to illustrate how models are specified. We provide R code in the text (and supplement) to facilitate implementation as R is free to use, and all code and data are also provided in our online repository. We focus on code to make the conceptual importance of random effects clearer to readers who may be less familiar with mathematical notation and matrix algebra.

In the example below, we have a single within-subject variable (`W1`), so we need to account for within-subject clustering of the data. Including a random intercept for each subject (`1|subject`) removes between-subjects variance from the outcome (`DV`), so that our independent variable is explaining within-subject variance after between-subjects variance has been accounted.

Code Snippet 1

```
# Note that R uses the '#' symbol to indicate programmer comments

example_model <- lmer(DV~
                        # Fixed-effects
                        1+ W1 +
                        # Random-effects
                        (1|subject),
                        data=DATA, REML=FALSE)

# Note that DV is the name of a continuous outcome, W1 is a factorial within-
# subject variable, and subject is a factorial subject identifier.
```

In the text, we indicate variables in a dataset with `code` markdown notation. For instance, we might generally speak of a "subject" participating in an experiment, but `subject` refers to a vector of subject identifiers (e.g., a column of "s1", "s2", as shown in Code Snippet 2). We also specify variables that are categorical factors versus those that are continuous predictors. For instance, when analyzing data from a randomized controlled trial, `condition` could be a vector of labels (e.g., "Experimental" versus "Control") whereas `time` could be a vector of times when data were collected (e.g., 0, 10, 35 days).

When presenting model outputs, we show F-tests that use Satterthwaite's approximation (Kuznetsova et al., 2017; Satterthwaite, 1941) to obtain degrees of freedom and p-values which will be of interest to most readers (Benjamin & Berger, 2019; Goodman, 2019; but also see Wasserstein et al., 2019). We use Satterthwaite's approximation for its computational efficiency and to obtain degrees of freedom that we can compare to factorial ANOVA results (Lawrence, 2016). There are other methods available such as the Kenward-Roger approximation (Kenward & Roger, 1997) or model comparison approaches (Judd et al., 2017; Long, 2012), but a detailed discussion of all of these approaches is beyond our scope. Similarly, there are methods like "sandwich" estimation (Gurka et al., 2011; Kauermann & Carroll, 2001) and bootstrapping (Efron & Tibshirani, 1993; Mason et al., 2021), which are useful for making inferences robust to violations of parametric assumptions, but a detailed explanation of those methods is beyond our focus on random effects. Finally, we do not use R^2 measures of effect-size as a direct R^2 calculation is not available in mixed-effect models, but see Johnson (2014) & Nakagawa & Schielzeth (2012) for pseudo-R^2 measures.

Defining Fixed and Random Effects

There are different definitions of fixed and random effects in the literature (B. M. Bolker et al., 2009; Gelman, 2005; Snijders & Bosker, 2012), but our definitions are consistent with the frequentist mixed-effect estimation employed in `lme4` (Bates, Machler, et al., 2015; Kuznetsova et al., 2017) and many domains of research (B. M. Bolker et al., 2009; Long, 2012). *Fixed effects* are those effects whose levels are fully represented in the data and beyond which we do not want to generalize. *Random effects* are those sampled from a larger population and from which we do want to generalize. For instance, in the code below, if we measured subjects' heart rate at rest and again following a bout of exercise, then `condition` (at rest or following exercise) would be a fixed effect. We are only interested in the difference between that specific type/duration of exercise and rest. In contrast, `subject` would be a random effect, because we have a random sample of subjects, and we want to generalize from that sample back to the larger population. Programmatically, that could be written:

Code Snippet 2

```
example_mod <- lmer(heart_rate~
                        # Fixed-effects
                        1+ condition +
                        # Random-effects
                        (1|subject),
                        data=DATA, REML=FALSE)

# Based on the data below:
> DATA
  subject condition heart_rate
1      s1       ctl         60
2      s1        ex         72
3      s2       ctl         42
4      s2        ex         50
```

Formally, the fixed and random effects in a regression model are written as a set of vectors and matrices:

Equation 1

$$y = X\beta + Z\gamma + \epsilon$$

where y is the vector of scores on the dependent variable for each person on each occasion it is measured, X is the design matrix for the fixed effects, β is a vector of coefficients, Z is a design matrix for the random-effects, γ is a vector of random deviates, and finally ϵ is a vector of random errors.

Although written in matrix notation, all we are really saying is that “the data = the model + the error” (Judd et al., 2017). However, our model has two components to it: the structural part, represented by the fixed-effects, $X\beta$, and the stochastic part, represented by the random-effects, $Z\gamma$ (Singer & Willett, 2003).

In this heart rate example, there were only 2 subjects (S1 and S2) measured in each of the 2 conditions: during a controlled rest, “ctl”, and following exercise, “ex”. As shown in Figure 2, we have one subject with a higher average heart rate (S1 = 60, 72) and one subject with a lower-than-average heart rate (S2 = 42, 50), in each of the conditions. Our random intercept of `(1|subject)` accounts for average between subject differences in y with the design matrix Z. This design matrix spreads out our random-effects, γ, into appropriate partitions for each subject; that is, every observation for S1 gets +9.9 and every observation for S2 gets -9.9. We can then add this result to the fixed effect estimates from the model. The fixed effect design matrix, X, indicates that there is a constant intercept (column of 1’s) that gets further modified by a slope in every other instance (column of 0’s, at the first observation, and 1’s, at the second observation). The specific values for this intercept and slope come from the fixed-effects coefficients, β. Thus, our model boils down to a regression equation for each person (i.e., 51+10(`condition`), where condition = 0 is rest and = 1 is exercise) with a unique deviate added to observations coming from the same subject (i.e., -9.9 or 9.9). Critically, accounting for the statistical dependency between observations with the appropriate random effects allows us to assume independent errors for each observation (estimated by the residuals, ϵ’s) and thus appropriately test hypotheses about the fixed effects.

Nested and Crossed Random Effects

Readers entering the world of mixed-effect models are likely to have heard of random-effects being either nested or crossed. *Nested* refers to situations where levels of one factor are represented at only one level of another factor. Such nesting occurs in fully hierarchical models (Raudenbush & Bryk, 2002; Singer & Willett, 2003). For instance, children are nested within families, students might be nested within classrooms, and classrooms may be nested within schools. Note that nesting is a product of the research design/environment in which a study is conducted and may not be immediately apparent in how variables are coded. Figure 3A shows an example of unambiguously coded nested factors: `students`

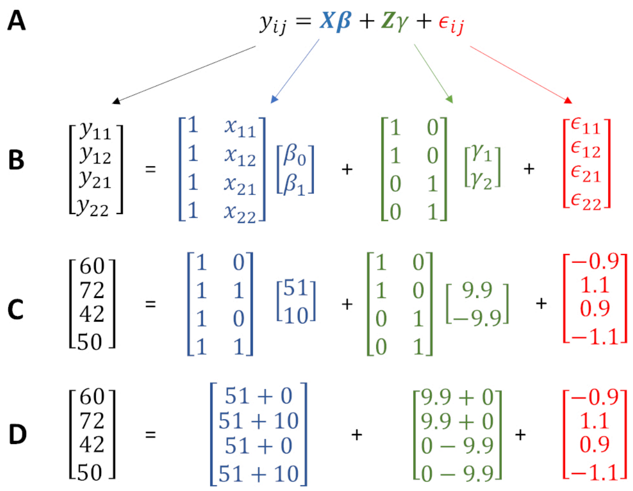
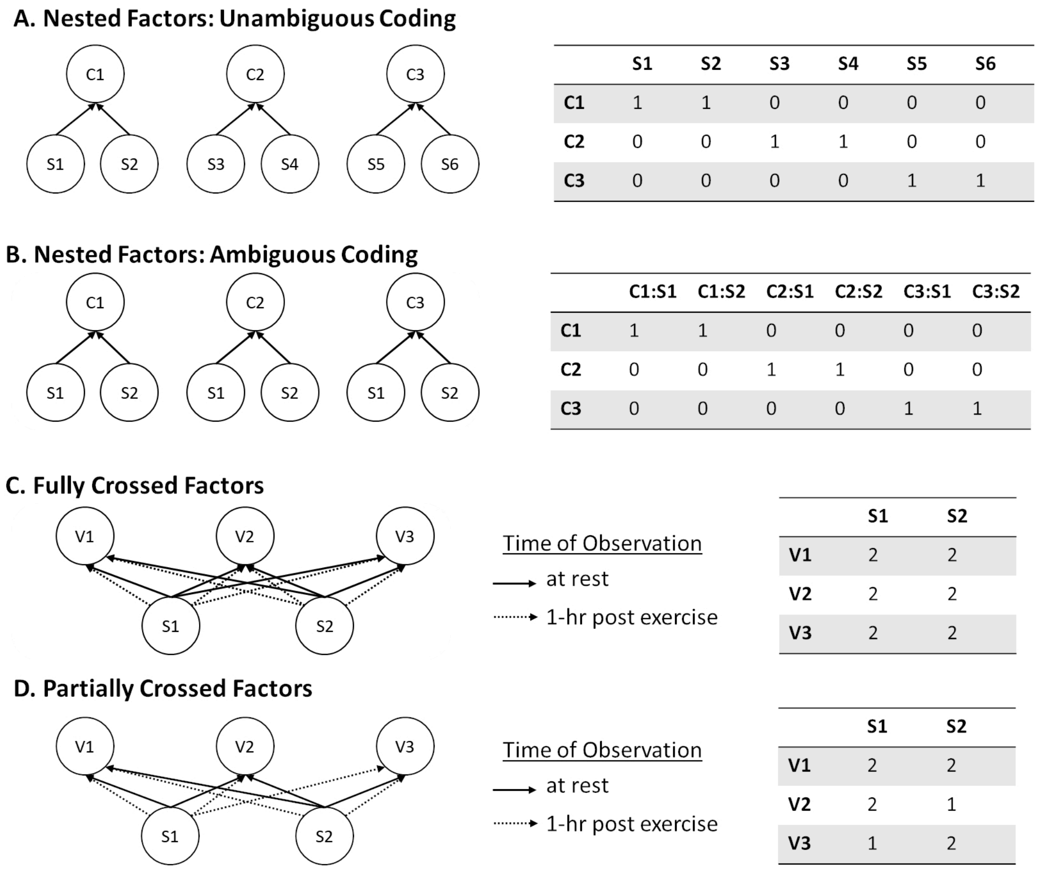

A
$y_{ij} = X\beta + Z\gamma + \epsilon_{ij}$
B
$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{21} \\ 1 & x_{22} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \end{bmatrix}$
C
$\begin{bmatrix} 60 \\ 72 \\ 42 \\ 50 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 51 \\ 10 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 9.9 \\ -9.9 \end{bmatrix} + \begin{bmatrix} -0.9 \\ 1.1 \\ 0.9 \\ -1.1 \end{bmatrix}$
D
$\begin{bmatrix} 60 \\ 72 \\ 42 \\ 50 \end{bmatrix} = \begin{bmatrix} 51+0 \\ 51+10 \\ 51+0 \\ 51+10 \end{bmatrix} + \begin{bmatrix} 9.9+0 \\ 9.9+0 \\ 0-9.9 \\ 0-9.9 \end{bmatrix} + \begin{bmatrix} -0.9 \\ 1.1 \\ 0.9 \\ -1.1 \end{bmatrix}$

Figure 2. Illustration of matrix notation for a mixed-effect model (A), which can be expanded into more detailed notation for our heart rate example with two observations and two subjects (B). Given the real values and the values for the design matrices, X and Z, (i.e., these matrices are provided by the researcher, not estimated from the data) one can estimate the values of the fixed- and random-effects, β and γ (C). We can then combine the fixed effects and the random effects with their respective design matrices, to get one estimate from our model (fixed + random effects) for each subject in each condition (D). The difference between the model's prediction and the actual value (y's) is the residual-error term (ϵ's).

(S1-6) are nested within different classrooms (C1-3) with no student belonging to multiple classrooms. As seen in the cross tabulation, each student has only one observation in only one of the `classrooms`. Figure 3B shows an example of an ambiguously coded nested factor. The same two levels of student (S1-2) show up at each level of each level of classroom (C1-3). However, student S1 in class C1 is not the same student as S1 in C2 or S1 in C3. Although it is best practice to code unambiguously, to make the nesting clear, we can concatenate the classroom and student labels (`classroom:student`) to get a unique identifier (e.g., C1:S2). The cross tabulation then shows the correct nesting: one observation from each student in only one class. If we did not concatenate these labels, we would end up with the wrong cross tabulation with 3-observations each for S1 and S2, one in each classroom.

Crossed refers to situations when different levels of one factor are represented at different levels of another factor. Two variables are said to be *fully crossed* if all levels of one factor are represented at all levels of the other factor. For instance, let's say that two subjects (S1-2) had to provide responses to video stimuli (V1-3) in our two different conditions: at rest and 1-hr post exercise. Thus, in a balanced design each subject has two observations for every video, one for each condition. Our focus is on the fixed effect of condition, but our randomly selected subjects and video stimuli need to be accounted for as random effects. If there are no missing data, we would get the fully crossed design shown in Figure 2C; each subject has three observations, one for each video for which a response was measured. If there are some missing data, these factors might become *partially crossed* as in Figure 2D (S1 is missing an observation for V3, S2 is missing an observation for V2).

Using `lme4` in R, researchers have several different options for inputting nested and crossed random effects as illustrated in Code Snippet 3.



A. Nested Factors: Unambiguous Coding

	S1	S2	S3	S4	S5	S6
C1	1	1	0	0	0	0
C2	0	0	1	1	0	0
C3	0	0	0	0	1	1

B. Nested Factors: Ambiguous Coding

	C1:S1	C1:S2	C2:S1	C2:S2	C3:S1	C3:S2
C1	1	1	0	0	0	0
C2	0	0	1	1	0	0
C3	0	0	0	0	1	1

C. Fully Crossed Factors

	S1	S2
V1	2	2
V2	2	2
V3	2	2

D. Partially Crossed Factors

	S1	S2
V1	2	2
V2	2	1
V3	1	2

Figure 3. Schematic illustration of nested (A/B) and crossed (C/D) factors with their cross-tabulations on the right. C1-3 refer to levels of the categorical factor C. S1-2 refer to levels of the categorical factor S. V1-3 refer to levels of the categorical factor V. In the nested case, we could think about S as students and C and classrooms (different students are in each classroom). In the crossed case, we have an example with S as subjects and V as video stimuli that were watched at two different phases of an experiment (at rest and 1-hr post exercise), yielding two observations for each subject per video (panel C), or one to two observations if some subjects/videos are missing data (panel D).

Code Snippet 3

```
# For students nested within classrooms we could use
(1|classroom) + (1|classroom:student)

# Or equivalent short cut using a slash "/":
(1|classroom/student)
# this syntax will work for ambiguous and unambiguous coding, see Figure 2

# If you have **unambiguously** coded nested factors, you can also use:
(1|classroom) + (1|student)


#For fully or partially crossed random effects, you would use:
(1|subject) + (1|video)
```

If you have ambiguously coded nested factors, that means you would need to specify `(1|classroom) + (1|classroom:student)`. Thus, there is a random intercept for each classroom and for each unique student (`classroom:student` concatenates the levels of Classroom, n=3, and Student, n=2, making

6 unique labels). `lme4` also provides a shorthand way of representing this using the ' / ' operator, `(1|classroom/student)`. If you have unambiguously coded data, you could instead use `(1|classroom) + (1|student)` and get the same result. This code is equivalent because in unambiguous data, student has six levels (S1-S6) and in ambiguous data, the concatenation `classroom:student` also has six levels ('C1:S1' to 'C3:S2').

For fully and partially crossed random effects, one would use the syntax `(1|subject) + (1|video)`. Note that this syntax has the same form as the unambiguously coded nested effects: `(1|classroom) + (1|student)`. This similarity helps to illustrate that the nesting (or crossing) of the random effect is not a property of the model, but a property of the experimental design. The model does not actually "know" which factors are crossed and which are nested, rather it depends on how those data are coded and selection of the appropriate syntax to suit the experimental design. Critically, a model will likely converge and produce an output of coefficients with any of these random effects, and no warning that a user has the wrong random effects. This highlights the importance of researchers thinking critically about how their data are coded, then selecting the appropriate random-effects syntax.

Longitudinal Designs

Random Slopes and Intercepts

To illustrate some of the modeling concerns with longitudinal data, let us consider the hypothetical study depicted in Figure 4A. This plot shows data from three different groups of participants who had different grades of spinal cord injury (high tetraplegia, cervical vertebral levels 1-4; low tetraplegia, cervical vertebral levels 5-8; or paraplegia, vertebral levels thoracic 1- sacral 5) and reported their functioning in the activities of daily living at approximately monthly intervals for 1.5 years as they underwent physical and occupational therapy. Functioning is an umbrella term for how body structures/functions (e.g., level of spinal cord injury) interact with personal and environmental factors (e.g., accessibility), to shape activity limitations and participation restrictions (Imrie, 2004; Üstun et al., 2003). Functioning in this case is the opposite of disability. For ease of interpretation, we constructed a fictitious measure of functioning on which scores range from 0 (very high disability) to 100 (very high function) and has interval/ratio properties. (Health science researchers might be more familiar with this construct as "independence", but we use "functioning" to avoid confusion with the statistical meaning of independence.)

As shown in Figure 4A, these data will likely violate the assumptions of the ordinary least squares regression model in which errors are assumed to be independent and identically distributed. This violation can be seen most clearly when we consider that these residuals belong not only to each time point but are also unique to each participant. Consider the lowest scoring participant in the C1-4 group in the left panel of Figure 4A. In a very simple model predicting the mean value for everyone (i.e., intercept only model), this participant is going to have negative residuals (i.e., the difference between the model's predictions and the observed value). In contrast, a high scoring participant will have mostly positive residuals. The exact value of the residuals will depend on the model, but in general we have two sources of dependence (scores from the same person tend to be more similar and scores from neighboring time points tend to be more similar) that will lead to correlations in the residuals (negatives with negatives, positives with positives), violating the assumption of independence. Violating this assumption is not trivial and can lead to considerable bias in the results of hypothesis tests (Dunlop, 1994; Sainani, 2010).

To account for this statistical dependence, the first step is to recognize that time points (`time`) are nested within subjects (`subject`) in this design. To do this, we can add a random intercept of subject to our model, as shown mathematically,

Equation 2A

$$y_{ij} = \beta_0 + \gamma_{0i} + \epsilon_{ij}$$

where:

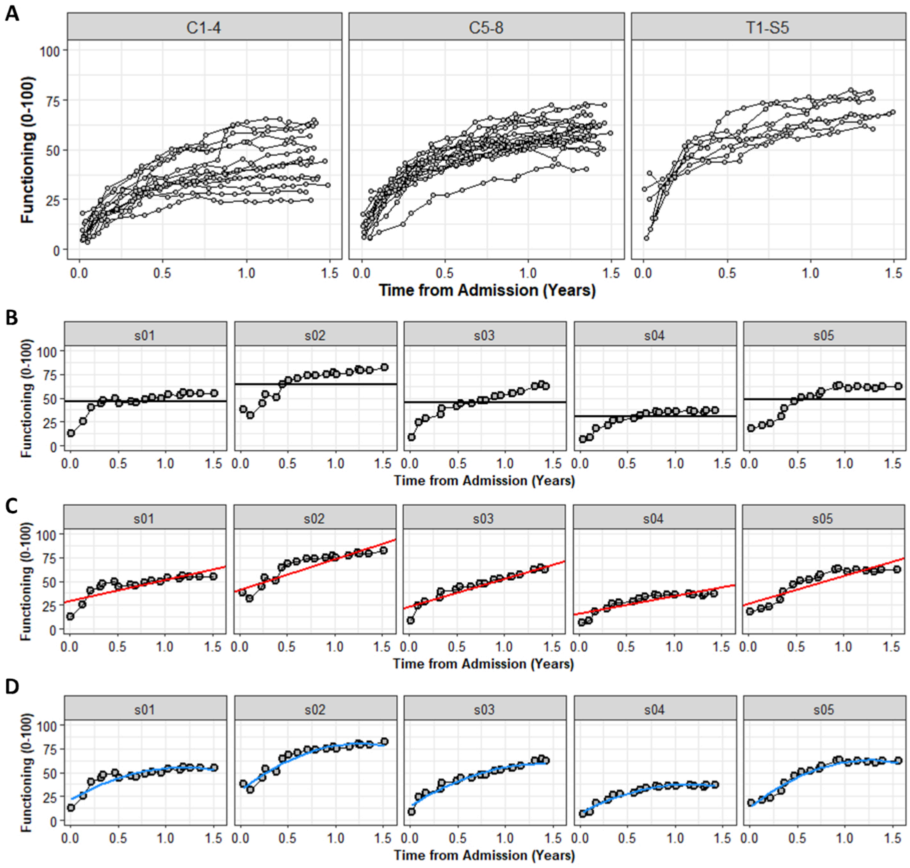


Figure 4. (A) Hypothetical functioning scores by time from admission for three different groups of participants with spinal cord injury. (B) Predicted values from a *random-intercepts model* (shown as black lines) plotted over the data for the first five participants. (C) Predicted values from a *linear random-slopes* model (shown as red lines) plotted over the data for the first five participants. (D) Predicted values from a *quadratic random-slopes* model (shown as blue lines) plotted over the data for the first five participants.

Equation 2B

$$\gamma_{0i} \sim N(0, \sigma_0^2)$$

or programmatically:

Code Snippet 4

```
raneff_int <- lmer(functioning~
                        # Fixed effects
                        1+
                        # Random effects
                        (1| subject),
                        data=DATA, REML=FALSE)
```

The predictions of this *random-intercepts model* are shown in Figure 4B for the first five subjects. Each subject gets their own unique intercept ($\beta_0 + \gamma_{0i}$), but there are large residuals not explained by the model (ϵ_{ij}). These residuals are the difference between each dot (real data) and the line (model estimate) in Figure 4B. By adding additional fixed and random effects to the model, we can reduce error at the level of the participant (i's) and each time point (j's).

To improve this model, we can estimate a unique trajectory for each subject by adding a fixed and random effect of time. This is often referred to as a *random-slopes model*, shown mathematically:

Equation 3A

$$y_{ij} = \beta_0 + \beta_1 \cdot (time_{ij}) + \gamma_{0i} + \gamma_{1i} \cdot (time_{ij}) + \epsilon_{ij}$$

where:

Equation 3B

$$\begin{bmatrix} \gamma_{0i} \\ \gamma_{1i} \end{bmatrix} = N \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}$$

Code Snippet 5

```
raneff_lin_slope <- lmer(functioning~
                           # Fixed effects
                           1+time+
                           # Random effects
                           (1+time| subject),
                           data=DATA, REML=FALSE)
```

The predictions of this linear random-slopes model can be seen in Figure 4C. Now, not only does each subject have a unique intercept ($\beta_0 + \gamma_{0i}$), but also a unique slope ($\beta_1 + \gamma_{1i}$), which greatly reduces the amount of residual error (ϵ_{ij}). Note also that one does not need to add these fixed and random effects in one step (Long, 2012; Raudenbush & Bryk, 2002; Singer & Willett, 2003). A fixed-effect could be added alone, which assumes that all participants changed at about the same rate, or a random-effect could be added alone, but that would assume the average effect of time is 0 across subjects. Importantly, having random-intercepts and random-slopes in the model also means that we must account for their covariance (σ_{01} in Eq. 3B). In applied modeling, it is common to find that individual differences in slopes and intercepts will be correlated: either positively because individuals with higher intercepts will show greater change over time (i.e., “the rich get richer”) or negatively correlated because individuals with lower intercepts will show greater change over time (i.e., “more room to grow” due to floor/ceiling effects).

Finally, we might want to account for the nonlinearity that we observe in the data. An intuitive way to account for this nonlinearity that is a direct extension of the linear model is to add a quadratic effect of time (i.e., `time^2`) to the model. Adding both a fixed- and random-quadratic effect of time yields:

Equation 4A

$$y_{ij} = \beta_0 + \beta_1 \cdot (time_{ij}) + \beta_2 \cdot (time_{ij}^2) + \gamma_{0i} + \gamma_{1i} \cdot (time_{ij}) + \gamma_{2i} \cdot (time_{ij}^2) + \epsilon_{ij}$$

where:

Equation 4B

$$\begin{bmatrix} \gamma_{0i} \\ \gamma_{1i} \\ \gamma_{2i} \end{bmatrix} = N \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

or programmatically:

Code Snippet 6

```
raneff_quad_slope <- lmer(functioning~
                         # Fixed effects
                         1+time+I(time^2)+
                         # Random effects
                         (1+time+I(time^2)| subject),
                         data=DATA, REML=FALSE)
```

In this model, as shown in Figure 4D, each participant not only has a unique starting point (the random intercept) and a unique initial slope at the intercept (the random linear slope), but also a unique curvature as the trajectory changes over time (the random quadratic slope).

Importantly, however, our model becomes much more complicated as we add random-slope parameters. Note that the addition of quadratic random-slope parameter did not merely add one parameter to our model, it actually added three: one for the variance of the quadratic slope (σ_2^2), one for the covariance of intercept and the quadratic slope ($\sigma_0 2$), and one for the covariance of the linear and quadratic slopes ($\sigma_1 2$), as shown in Eq. 4B. Thus, adding fixed- and random-effects of polynomials can improve the fit of our model, but we need to be thoughtful when doing so. Additionally, researchers need to consider the centering of the time variable when adding polynomials to the model. If time is centered on the first observation (i.e., `time` $= 0$ reflects the first observation) then higher polynomials will often be highly correlated, which can create convergence issues for the model. For detailed accounts of how to build these polynomial models, we refer readers to more thorough sources that explain how to interpret these effects, test the variance explained by each random-effect, and compare models with different random-effects in order to choose the most parsimonious model that still fits the data appropriately (Long, 2012; Singer & Willett, 2003).

Random-Effects in Non-Linear Models

The quadratic random slopes model above is what we refer to as a *curvilinear model.* That is, the model ultimately predicts a nonlinear trajectory for each person, but the model itself is linear in its components: β_0 is simply added to $\beta_1 \cdot time_i$ and added to $\beta_2 \cdot time_i^2$ as in Eq. 4A. Forcing this nonlinearity via polynomials has strengths and weaknesses. Strengths include the ease of interpreting polynomials (e.g., the average reader can more easily conceive of `time^2` compared to the log of `time` or `time` exponentiated) and they often correspond to hypotheses that researchers might have (e.g., the typical researcher might be interested in *any* evidence of curvature, rather than a specific power law or exponential law). However, a major weakness is that few biological or social systems follow a quadratic function (i.e., being u- or n-shaped and symmetric about a single inflection point). For instance, in Figure 4D, it is likely that participants were approaching an asymptote that could be better captured by a truly nonlinear model such as an exponential or logistic model.

A detailed discussion of nonlinear models lies outside of our focus on random effects, but see Chapter 6 of Singer & Willett (2003), and Pinheiro & Bates (2006). To summarize however, (1) regardless of the form the nonlinear model takes, there is the possibility to include random effects for each parameter, and that (2) adding random effects can often lead to issues of over-fitting, so the random effects included in the final model should be justified through a combination of theoretical rationale and empirical model fit.

Here, we present an illustrative example of a negative exponential model. First, let us consider a three-parameter exponential function:

Equation 5A

$$y_{ij} = [\beta_0 + \gamma_{0i}] + [\beta_1 + \gamma_{1i}] \cdot e^{-(\beta_2+\gamma_{2i})\cdot time_{ij}} + \epsilon_{ij}$$

where:

Equation 5B

$$\begin{bmatrix}\gamma_{0i}\\ \gamma_{1i}\\ \gamma_{2i}\end{bmatrix} = N\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\sigma_0^2 & \sigma_{01} & \sigma_{02}\\ \sigma_{01} & \sigma_1^2 & \sigma_{12}\\ \sigma_{02} & \sigma_{12} & \sigma_2^2\end{bmatrix}$$

Our model now consists of fixed effects, β_0,β_1, and β_2, which apply to the whole sample, shown in Figure 5A. However, we also have unique asymptotes $(\beta_0 + \gamma_0 i)$, distance to the psuedo-intercept $(\beta_1 + \gamma_1 i)$, and rate parameters $(\beta_2 + \gamma_2 i)$ for each subject which are the sum of the fixed effect and the random deviate for that subject. These coefficients are shown for the first four subjects in Figure 5B.

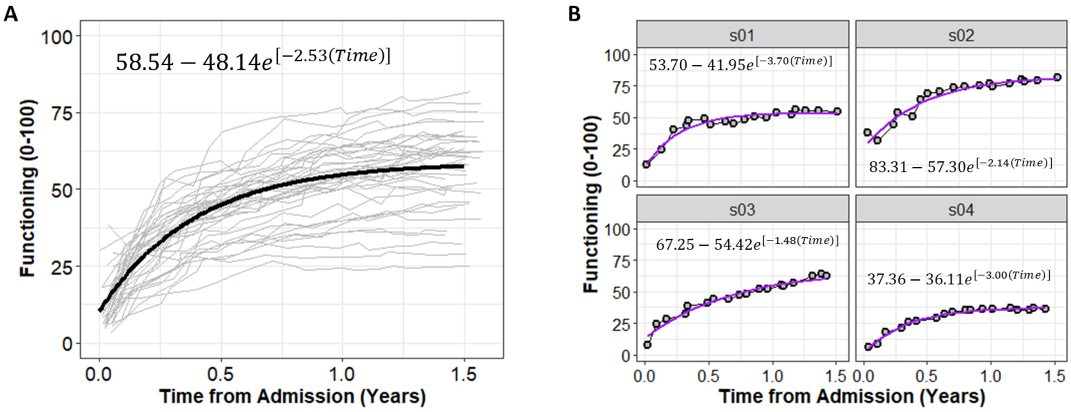


Figure 5. (A) Modeling the same hypothetical spinal cord injury data using a negative exponential function. Grey lines show the data for individual subjects. The black line shows the best fitting negative exponential model, with the fixed effects displayed in the panel. (B) Individual data (as points) and estimated trajectories (as purple lines) for the first four subjects. Equations show the coefficients for each participant (the group-level fixed effect from A, plus the random deviate for each participant).

Programmatically, we can also see how this model would be fit using `nlme` (Pinheiro et al., 2022) in Code Snippet 7. A worked example of this non-linear model is provided in the online repository (as there are some important programming considerations for `nlme` versus `lme4`).

Code Snippet 7

```
set.seed(100)
neg_exp_rand_mod <- nlme(functioning ~ b_0i +
                           (b_1i)*(exp(b_2i * time)),
                         data = DATA,
                         fixed = b_0i + b_1i + b_2i ~ 1,
                         random = list(b_0i ~ 1, b_1i ~ 1, b_2i ~1),
                         groups = ~ subject,
                         start = c(80, -70, -1),
                         na.action = na.omit)
```

Repeated Measures Factorial Designs

In this section, we will consider situations in which we have categorical repeated measures, not continuous longitudinal data. For instance, a researcher might have a design with two within-subject factors of `condition` (conditions of rest, immediately following exercise, and following a 30-min delay) and `altitude` (laboratory simulated low- versus high-altitude environments). Although these measurements were taken at different times of day, a continuous measure of time is not what we want to model.

Instead, we want to model the categorical difference between the two different sets of repeated measures: condition and altitude as shown in Figure 6A. Researchers who use these types of within-subject designs have historically analyzed their data using repeated-measures ANOVA, in which all factors vary within a subject. In our experience, many researchers want to apply the same conceptual analysis but utilize some of the strengths of a mixed-effect model (e.g., avoiding listwise deletion of missing data). But what is the appropriate random-effects structure that will make the model analogous to the repeated measures ANOVA? Consider five different options in Code Snippet 8:

Code Snippet 8

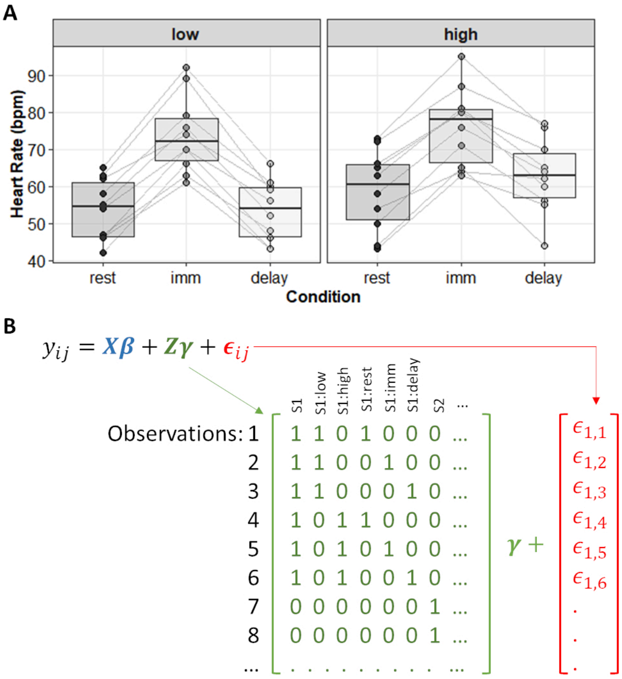
```
# Head of the data file showing the first 12 observations (subject S1 and S2)

   subject altitude condition heart_rate
1       S1      low      rest         47
2       S1      low       imm         63
3       S1      low     delay         48
4       S1     high      rest         54
5       S1     high       imm         64
6       S1     high     delay         62
7       S2      low      rest         62
8       S2      low       imm         74
9       S2      low     delay         60
10      S2     high      rest         65
11      S2     high       imm         80
12      S2     high     delay         70
# …

# Option A: Random Intercept of Subject Only
heart_rate ~ 1 + condition*altitude +
             (1|subject)


# Option B: Random Intercepts for All Factors
heart_rate ~ 1 + condition*altitude +
             (1|subject)+
             (1|condition)+
             (1|altitude)


# Option C: Random Intercepts of Subject:Repeated Measures
heart_rate ~ 1 + condition*altitude +
             (1|subject) +
             (1|subject:condition) +
             (1|subject:altitude)


# Option D: Random slopes of condition and altitude within each subject
heart_rate ~ 1 + condition*altitude +
             (1+condition+altitude|subject)


# Option E: Random slopes including the interaction term
heart_rate ~ 1 + condition*altitude +
             (condition*altitude|subject)
```

When we analyze these data using a repeated measures ANOVA, we obtain main effects of Altitude, F(1,9) = 58.36, Condition, F(2,18)=142.15, and the Altitude × Condition interaction, F(2,18)=8.63. We can use these values as a benchmark to evaluate the performance of our mixed-effect models. Using these values as our "gold standard" assumes that the factorial ANOVA is the model that researchers want given their design. We acknowledge that this may not always be the case, but we think it is generally best to choose random effects based on the study design and given that most researchers would analyze data from this study with a two-way repeated-measures ANOVA, it is reasonable to use that benchmark.

Before we can evaluate these different random effects structures, it is important to briefly explain two different methods of estimation: Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML). Without getting into the details of how these estimators actually work (Bates, Machler, et al., 2015; Pinheiro & Bates, 2006; Singer & Willett, 2003) a key difference is that ML will base estimates on all of the available information (i.e., fixed and random effects) but does this as an iterative procedure. As a conceptual approximation, for each iteration of ML the fixed components are estimated first, and then random components estimated second given those fixed effects. Mathematically, this happens simultaneously, but conceptually we can think about the fixed effects being needed first, as the sample mean needs to be calculated to then calculate the sample variance. As with the calculation of the sample variance, this means that variances in ML tend to be under-estimated. However, this bias in ML versus REML should only substantially affect the estimation of the variance when the number of observations, n, is small relative to the parameters being estimated, p. Readers are likely familiar with this issue when calculating the variance of a sample, where we divide the sum of squared errors by $n-1$ (analogous to REML) rather than simply n (analogous to ML). Thus, the de-biasing effect of dividing by $n-p$ relative to just dividing by n depends on the values of both n and p. When the number of observations is much larger than the number of parameters ($n >> p$), then the choice between REML and ML will usually have a trivial effect on the variance components. Importantly, however, REML does not correct this bias simply by changing the denominator of the variance components, but by using a fundamentally different algorithm for estimation.

In REML, the outcome is effectively transformed into residuals, variance components are then iteratively estimated, and then the fixed effects are estimated using generalized least squares (for an accessible discussion see McNeish (2017)). Thus, unlike ML where we simultaneously estimate the fixed and random effects in every iteration, with REML the fixed effects can essentially be treated as 0 at first because we are estimating residuals, and partitioning that variance based on the best fitting values of the random effects. In the final step of REML, we thus already have a variance/covariance matrix for the random effects and can estimate the best fitting fixed effects given that matrix.

These computational differences mean that REML and ML have complementary strengths and weakness. For instance, if you wanted to compare models with different fixed effects to ascertain the best fitting model, then you need to use ML because it is agnostic to the degrees of freedom lost to the different fixed effects in the different models. However, ML is also prone to under-estimate the variance components, especially in small samples. Thus, if there is no model comparison (e.g., you are a fitting a single mixed model with random effects determined by the design) or if you are comparing models with identical fixed effects (e.g., trying to determine only the preferred random effect structure), then REML will lead to less biased estimation of the variance components than ML.



$$y_{ij} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \epsilon_{ij}$$

$$\begin{array}{c|cccccccc|}
 & S1 & S1{:}low & S1{:}high & S1{:}rest & S1{:}imm & S1{:}delay & S2 & \ldots \\
\hline
1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & \ldots \\
2 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & \ldots \\
3 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & \ldots \\
4 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & \ldots \\
5 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & \ldots \\
6 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & \ldots \\
7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \ldots \\
8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \ldots \\
\ldots & . & . & . & . & . & . & . & \ldots
\end{array}
\gamma + \begin{bmatrix} \epsilon_{1,1} \\ \epsilon_{1,2} \\ \epsilon_{1,3} \\ \epsilon_{1,4} \\ \epsilon_{1,5} \\ \epsilon_{1,6} \\ . \\ . \\ . \end{bmatrix}$$

Figure 6. (A) Data from our hypothetical experiment in which the same subjects had their heart rate measured at simulated low and high altitude in the lab at three different time points: at rest, immediately following exercise, and following a 1-hr delay post exercise. (B) A truncated design matrix for the random-effects structure of the model. The first 8 observations (6 belonging to subject S1, 2 to subject S2) are shown in rows and the 'Subject:Factor' intercepts from the mixed model are shown in columns.

Table 1. Summary model output for a factorial repeated measures ANOVA and mixed-effects models using five different random-effects and two different methods of estimation (ML versus REML).

	Main Effect of Altitude		Main Effect of Condition		Altitude x Condition Interaction	
	(df1, df2)	F-value	(df1, df2)	F-value	(df1, df2)	F-value
Repeated Measures ANOVA						
ANOVA, Type III sums of squares	(1, 9)	58.36	(2, 18)	142.15	(2, 18)	8.63
Random Effects - REML						
Option A	(1, 45)	43.13	(2, 45)	199.25	(2, 45)	6.29
Option B	(1, 45)	0.5	(2, 45)	1.52	(2, 45)	6.29
Option C	(1, 9)	58.35	(2, 18)	142.15	(2, 18)	8.63
Option D[A]	(1, 12)	43.88	(2, 9)	123.51	(2, 27)	10.51
Option E[B]	N/A	N/A	N/A	N/A	N/A	N/A
Random Effects - ML						
Option A	(1, 50)	47.92	(2, 50)	221.39	(2, 50)	6.99
Option B[C]	(1,50)	47.92	(2, 50)	221.39	(2, 50)	6.99
Option C	(1, 10)	64.84	(2, 20)	157.95	(2, 20)	9.59
Option D[A]	(1, 13)	48.77	(2, 10)	137.26	(2, 30)	11.68
Option E[B]	N/A	N/A	N/A	N/A	N/A	N/A

Notes: Readers can see all the underlying code, including the standard ANOVA functions in the online repository. Df1 refers to the numerator degrees of freedom and df2 refers to the denominator degrees of freedom for the calculation of the F-statistic. REML = Restricted maximum likelihood estimation; ML = Maximum Likelihood estimation.

[A] returns warning ”boundary (singular) fit: see help(’isSingular’)”, correlations between random-effects approach 1, but there is no zero variance for every random effect.

[B] Leads to an error, ”Error: number of observations (=60) <= number of random effects (=60) for term (condition * altitude | subject); the random-effects parameters and the residual variance (or scale parameter) are probably unidentifiable”. Without replicates, this model is over-specified and cannot be fit to these data.

[C] returns warning ”boundary (singular) fit: see help(’isSingular’)”, the variance for (1|altitude) and (1|condition) both go to zero, effectively leaving this the same as Option A when fit using maximum likelihood.

Option A: Random Intercept of Subject Only – A Poor Choice.

In our experience, many researchers select Option A, including only a random intercept of subject, `(1|subject)`. Including a random intercept for each subject is *necessary but not sufficient* to account for all the statistical dependencies in this within-subject design. This can be seen in Table 1, where we contrast the denominator degrees of freedom and F-values from a mixed-factorial ANOVA (Lawrence, 2016) and the five different mixed-effect models. Option A leads to greatly inflated denominator degrees of freedom for each of the effects; the denominator degrees of freedom are all 45, much larger than the corresponding 9 or 18 degrees of freedom in the repeated measures ANOVA. This spuriously increases statistical power (note the larger F-values for each test), in a loose sense making it seem as though we have more independent pieces of data than we do. Inspecting the degrees of freedom helps to quantify the conceptual problem with this random-effects structure: there a several "strata" in the data (Venables & Ripley, 1997) at which we could look at the effect of altitude and condition. With N=60 observations and only a random intercept of subject, we have essentially divided the data into two strata: a between subjects' stratum (n=10, df=9 after the grand mean is estimated) and within subjects' stratum (df=50 after between subject variance is removed). However, not all observations within a subject are independent of each other, as each subject has multiple observations at each altitude and condition.

Option B: Random Intercepts for All Factors – A Very Bad Choice.

To account for the fact that observations are clustered within subjects, altitudes, and conditions, we have often seen authors use a random-effects structure like Option B, treating altitude and condition as random effects. At first, these three random intercepts might seem reasonable, but there is a major conceptual flaw with this model: Our observations are not crossed by altitude and condition *overall*, they are only crossed by altitude and condition *within each subject*. The correct way to handle this is by nesting altitude and condition within each subject and accounting for those crossed factors at the within-subject level, which is precisely what we do in Option C. Before discussing that structure, however, let us dwell on Option B as an example of what not to do. As we will see, it is bad practice to enter a variable as both a fixed and random effect, because it is statistically non-sensical.

When estimated with REML, the main-effects for altitude and condition all but disappear with F(1, 45)=0.50 and F(2,45)=1.52, respectively. These F-values are greatly reduced because we have pulled out the variance due to `altitude` and `condition` in our random intercepts prior to the estimation of the fixed effects, leaving a trivial amount of variation to be explained. Recall that for REML, variance components are estimated from a residualized outcome, and fixed effects are then estimated using generalized least squares given the matrix of random effects. The interaction term remains similar to what we saw with Option A, F(2,45)=6.29, because that variance is not accounted for by the random intercepts.

When estimated with ML, the results of Option B are identical to Option A because now the fixed effects are iteratively estimated with the redundant random effects, which leaves no variance to be explained by the random intercepts of altitude and condition. The variance for both random intercepts is now estimated to be 0 (see footnotes for Table 1), and we are effectively left with a random intercept of `subject`.

These wildly varying estimates for altitude and condition help to illustrate how our different estimators work (i.e., REML vs. ML) and hopefully illustrate how incorrect Option B is. It is non-sensical to treat `altitude` as both a fixed effect and random effect, it needs to be one or the other. Note, how this is also different from longitudinal models where we estimated a fixed effect of `time`, but also had a random slope for `time` within each subject (i.e., `(1+time|subject)`). It is perfectly acceptable to estimate a fixed effect at the group-level, while accounting for individual differences in an effect through the random effects, discussed in Options C-E below.

Option C: Random Intercepts for Subject:Repeated Measures – A Good Choice.

Random effects Option C is a very good choice to appropriately account for the statistical dependencies in these data. In that model, the degrees of freedom and the F-values match the repeated measures ANOVA exactly using REML. These outputs match because not only are we accounting for the fact that datapoints come from the same subject, `(1|subject)`, but we also account for the fact that each subject can respond differently to each altitude, `(1|subject:altitude)`, and each condition, `(1|subject:condition)`. As shown in a truncated design matrix in Figure 6B, having a unique intercept for each subject:factor accounts for the fact that altitude and condition are crossed factors, within each subject. We have thus set up appropriate strata in the data at which we can evaluate our fixed effects (Venables & Ripley, 1997).

Note also that we cannot include an intercept of `(1|subject:condition:altitude)` because that would amount to a random deviate for every observation, leaving no residual error (i.e., such a three-way random intercept is confounded with the error term when we only have one observation person per condition, and `lme4` would return an error).

As shown in Table 1, if we use ML instead of REML the results no longer match the ANOVA precisely. The degrees of freedom are slightly increased, as would expected given the differences in estimation, and this in turn leads to slightly increased F-values. Thus, the choice of estimation method will have an effect when it comes to statistical inference, but if we fit these models in a large data set (where $n >> p$), the practical difference between REML and ML would be small.

It is important to note that this random-intercepts method gets more complicated with more within-subject factors. For instance, if we were to add a third within-subject factor to a fully factorial design (call it W3), then not only would this add an additional fixed effect, but it adds considerable statistical dependency that needs to be accounted for in the random effects. In that situation, observations come not only from each subject in each condition, but from each subject in each combination of the conditions, so we would need to add intercepts to account for the two-way interactions within each subject (e.g.,`(1|subject:condition:altitude) + (1|subject:altitude:W3) + (1|subject:condition:W3)`). It is also important to note that our heuristic for the degrees of freedom may start to breakdown in more complex designs and although the degrees of freedom will be similar to the ANOVA, they will no longer match exactly (e.g., with sparse random-effects when there may only be two-levels of a factor (Venables & Ripley, 1997)). To save on space, we present a detailed example of a three-way within-subject design in the online supplement.

Option D: Random Slopes of Condition and Altitude within in Subject – Also a Good Choice.

Another reasonable alternative would be to include random slopes for each factor within each subject as in Option D. Conceptually, this model accounts for the within-subjects nature of the design similar to Option C, but it approaches the problem in a very different way and with more complexity that may not always be desirable. First, recall that when fitting random slopes, we do not only estimate their variance, but also their covariance. This allows us to account for potential correlations between the effect of altitude and condition across subjects, which is good, but increases the number of parameters being estimated, which is potentially problematic.

To simplify the model, one could also constrain the random slopes to be independent (e.g., using "||" in place of "|" in `lme4` syntax, and see Frossard & Renaud (2019)). However, one would still run into a second issue which is that random slopes are actually a series of contrasts for categorical factors. That is, for `altitude` we only have two levels and therefore we need only one random slope to capture that difference. For `condition` we have three levels and therefore two random slopes are needed, which would also add more covariances if the model is not constrained. As the number of levels of the categorical factor/s increases, the model becomes substantially more complicated.

Third, and finally, when using this random-slopes approach, the user needs to cognizant of what kind of contrasts are being used (as with centering of continuous variables). By default, `R` will use treatment coding with the first level, alpha-numerically, treated as the reference group, but this may not always be desirable and will affect the estimates and interpretation of the model, particularly when interactions are present.

Option E: Random slopes including the interaction term – Only if you have replicates!

Estimating the variance around our effects is a very useful step towards understanding individual differences (e.g., how much does the effect of altitude differ between people?), so the random slopes structure in Option D is appealing. Ideally, we would like to take this a step further and get an estimate of how the interaction between altitude and condition differs between people as well, so we can add the random slopes that capture that interaction in Option E. However, we cannot actually estimate this model given the current data. With only one observation for each subject in each condition, if we try to fit a unique three-way interaction for every subject, we will have perfect prediction and no residual error (as with a random intercept of `(1|subject:condition:altitude)`). This can be seen in the footnotes under Table 1, we only have 60 observations and if we want random slopes for each factor and their interaction, then we end up with 60 random effects. As the error message says, "Without replicates, this model is over-specified and cannot be fit to these data".

Replicates, as the name implies, are repeated observations at the same level of the data. For instance, assume that rather than one average heart rate for each subject in each condition, we ran our subjects through our exercise protocol on two different occasions with a sufficient washout period in between exposures to reduce order effects. We would now have two observations for each subject in each combination of altitude and condition, yielding 120 observations. Having only two observations per person may not be enough to get very stable estimates of the variance for these within-subject effects, but it would give us the observations necessary to estimate a model using Option E. Thus, having replicates is often very desirable but comes with increased complexity and often implicit assumptions about which an author might not be aware. For instance, let's be generous and say that I have 10 observations for every combination of altitude and exercise. I could estimate a model using Option E, but that would assume that all 10 replicates are independent of each other. Depending on how those observations were obtained, independence might not be a tenable assumption. For instance, there are likely temporal order effects with neighboring observations being more similar. As a result, our model might need to become more complex, for instance estimating a covariance structure for the residuals rather than assuming that all errors are independent (Pinheiro et al., 2022; Pinheiro & Bates, 2006).

Designs with Multiple Sources of Variance

A major strength of mixed-effect models over traditional ANOVA is their ability to handle different sources of variance simultaneously. We can see these different sources of variance when we have multiple factors that are random samples. For instance, this can happen when we have multiple levels of nesting, such as time within patients and patients within hospitals. That is, we want to generalize from our sample of patients and our sample of hospitals back out to a theoretical population of patients and population of hospitals, but we also need to account for the statistical dependence of patients from the same hospital. This sort of hierarchical nesting is why many mixed-effect models are referred to as "multi-level models" (MLM's Snijders & Bosker (2012)) or "hierarchical linear models" (HLM's Raudenbush & Bryk (2002)). In the Variation in Multiple Nested Factors section below, we illustrate this kind of hierarchical nesting. We use the same hypothetical spinal cord injury data from the previous example of longitudinal data analysis. However, we now pretend that these patients were recruited and assessed at different study sites; so, we have time within subjects and subjects within sites. See Figure 7A.

We can also see different sources of variation when participants are repeatedly exposed to stimuli. In speech pathology and audiology, the stimuli might be a series of words that the participant is hearing or saying. The words are a random sample of possible words and therefore we want to account for the variance in our sample of words in the same way we account for the variance in our sample of subjects. In the Variation in Multiple Crossed Factors section below, we will use a linguistic example from speech recognition (Brown & Strand, 2019). These response time data have been modified from the original data for didactic purposes (Brown, 2021) and illustrate how we can account for variation due to both our sample of subjects and the sample of words. See Figure 7B/C.

In that experiment, participants listened to words in both an audio only modality in which a speaker was heard but not seen, and in an audiovisual modality in which the speaker was both seen and heard. Across all words were presented in both modalities, but each word was presented in only one modality for a given subject. The dependent variable was subjects' response time in a secondary task in which they had to decide if a tactile stimulus was presented for a short, moderate, or long duration. The response on this secondary task is thus an index of the cognitive demand of processing the word. If the response time is faster, we can infer that the primary task required fewer cognitive resources. If the response time is slower, we can infer that the primary task required more cognitive resources. The researchers hypothesized that participants would be slower to respond in the audiovisual condition because participants would be processing both the audio and visual information.

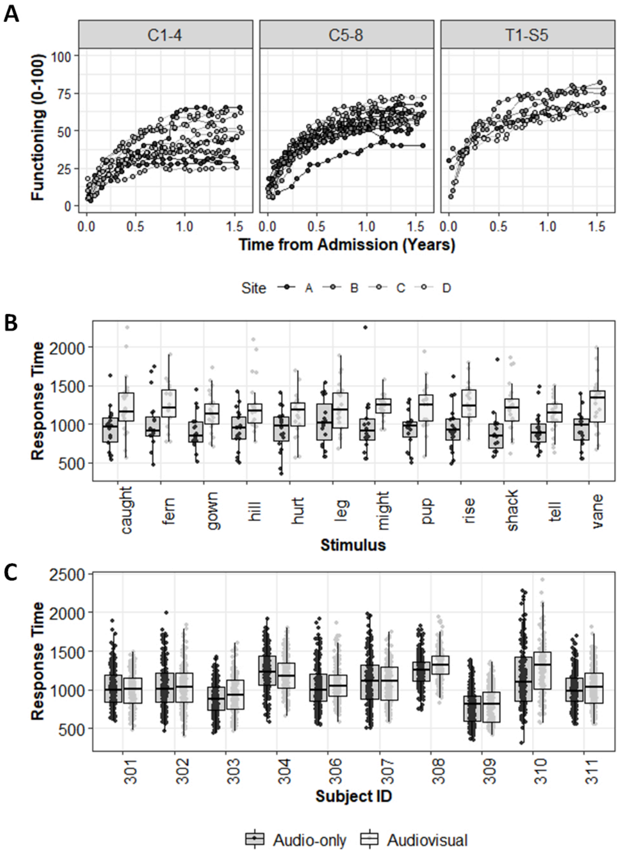


Figure 7. (A) Data from the hypothetical spinal cord injury study. Functioning in daily life is shown by time since admission and the study site at which the data were collected. Different study sites are shown in the fill colors, lines connect data points for individual subjects. (B/C) Data from the adapted linguistics example. Panel B shows response times as a function of the first ten words in the stimulus list (out of 543 total words). Points show the response time for individual subjects. Panel C shows response times as a function of the first ten subjects (out of 53 total subjects). Points show response times for individual words. In both B and C, points are color coded by the modality in which they were presented audio only, or in an audiovisual modality in which the speaker was seen.

Variation in Multiple Nested Factors (e.g., Subjects within Study Sites)

Returning to our hypothetical spinal cord injury data, in Figure 7A we can see multiple sources of statistical dependency. Time points are nested within individual subjects, but now we also have subjects nested within different research sites (A-D). We can account for this additional level of nesting in our `R` code by using a random intercept of study site. We also include a fixed effect of spinal cord injury grade (C1-4, C5-8, T1-S5) and *its* interactions with time:

Code Snippet 9

```
raneff_quad_site<-lmer(functioning~
                           # Fixed-effects
                           1+time*grade+I(time^2)*grade+
                           # Random-effects
                           (1+time|subject)+(1|site), data=DAT1, REML=FALSE)
```

Fitting this model yields the random and fixed effects shown in Table 2, and yields the warning message "`boundary (singular) fit: see ?isSingular`". Looking at the random-effects output in Table 2 we can see why: the variance for the `site` random-effect is estimated to be zero, which is a "boundary" because variances can only be between $[0, +\infty)$. Correlations are similarly bounded between $[-1, +1]$. In this case, it would be reasonable to drop the site random effect from the model, as there is no variance being reliably explained by the different study sites.

Table 2. Summary of the model output for spinal cord injury data, including a random intercept for the different study sites.

Random Effects				Fixed Effects			
Group	Name	SD	r	Effect	F-Value	df1, df2	p-value
Subject	Intercept	4.62		Grade	5,073.00	2, 50.7	<0.001
	Time	6.93	-0.08	Time	1,450.21	1, 140.4	<0.001
Site	Intercept	0.00		Time^2	1,104.46	1, 640.4	<0.001
Residual		3.44		Time x Grade	3.83	2, 141.1	0.024
				Time^2 x Grade	1.31	2, 640.5	0.271

Note that the number of observations = 720, subjects = 40, and sites = 4.

However, there are a few considerations for dropping this random effect from the model. First, we only have four study sites in this example, which is very small number of levels for the estimation of a random-effect (B. Bolker, 2023; Hodges, 2016). If we still wanted to account for differences between sites, we could include site as a fixed effect. If we include it as a fixed effect, we again see that there is very little variance explained by study site, F(3, 40.0)=0.97, p=0.414. So, in a practical sense, we can account for the variation due to study site in either format. Philosophically though, these two approaches are different. Treating study site as a fixed effect assumes that these are the only study sites in which we are interested. Treating study site as a random effect assumes sites are a random sample of possible sites. In either case, we should think about this lack of explained variance as a null result. That is, we have not proven that there is 0 variation between study sites, we have simply failed to find evidence of a difference between study sites – and the absence of evidence is not evidence of absence (60).

Variation in Multiple Crossed Factors (e.g., Variation in Subjects and Stimuli)

In our linguistic example, we must account for the fact we have multiple observations not only for each participant, but also for each stimulus. Note that unlike the longitudinal data, there is no reliable pattern to the data within a participant, because the order of the stimuli was randomized for each person. As such, we can account for this statistical dependence with random intercepts of `subject` and `stimulus`, and a fixed effect of the modality of the stimuli (`modality`).

Code Snippet 10

```
rand01 <- lmer(log(RT)~
                 # Fixed Effects
                 1+modality+
                 # Random Effects
                 (1|subject)+(1|stimulus),
               data=DAT2, REML=TRUE)
```

As shown in Table 3, there was relatively less variability due to stimuli (SD = 0.017) compared to subjects (SD=0.152). Accounting for these sources of variation, there was a statistically significant effect of modality, such that participants were generally slower to respond to the audiovisual stimuli compared to the audio-only stimuli, b=0.079, F(1,21440)=541.48, p<0.001.

Table 3. Summary of the model output for the linguistic data, including random intercepts of both participant and stimulus.

Random Effects			Fixed Effects			
Group	Name	SD	Effect	F-Value	df1, df2	p-value
Stimulus	Intercept	0.02	Modality	541.48	1, 21440	<0.001
Subject	Intercept	0.15				
Residual		0.25				

Note that the number of observations = 21,679, stimuli = 543, and subjects = 53.

However, we should take note of the denominator degrees of freedom for this F-test, which are 21,440. This makes sense given the large number of observations: 543 stimuli x 53 participants - missing data = 21,679 observations. However, this is much larger than what we would expect from a comparable ANOVA design if modality had been included as a within-subject factor: $df_2 = (N-1)(k-1) = (53-1)(2-1) = 52$. The reason that these degrees of freedom are so much larger is because we have numerous replicates within each subject. That is, for each subject in each modality, we have multiple observations from several different stimuli. In this situation, we have a few options. We could aggregate the data down to one observation for each subject in each modality. This would make our mixed model more like a traditional repeated measures ANOVA with a within-subject factor of modality. Indeed, if we averaged over stimuli and then ran the code below, we would obtain a main effect of modality with F(1,52)=45.94, p<0.001.

Code Snippet 11

```
DAT3 <- DAT2 %>% group_by(subject, modality) %>%
  summarize(RT = mean(RT))

agg_mod <- lmer(log(RT)~
                 # Fixed Effects
                 1+modality+
                 # Random Effects
                 (1|subject),
               data=DAT3, REML=TRUE)
```

However, aggregating over stimuli would remove all the variability due to the individual stimuli and increase the Type I error rate. Thus, if we want to estimate the variability due to the specific stimuli we chose, we need to retain the different observations for the different stimuli. Instead of aggregating, we could retain the random intercept of `stimulus`, and add a random slope for the `modality` within each subject as in Code Snippet 12.

Code Snippet 12

```
rand_slopes <- lmer(log(RT)~
                  # Fixed Effects
                  1+modality+
                  # Random Effects
                  (1+modality|subject)+(1|stimulus),
                data=DAT2, REML=TRUE)
```

We can see the results from this model in Table 4. By including a random intercept for stimuli and a random slope for modality within each subject, we can effectively estimate the variance due to the different stimuli (SD = 0.075) while reducing the degrees of freedom in the F-test, F(1, 51.9)=52.86, p<0.001. Conceptually, the reason the denominator degrees of freedom have shrunk from 21,440 to 51.9 is because we are no longer treating the modality effect as being estimated from 21,679 data points, but from 53 observations of the modality effect, one for each participant. Thus, if the model converges, including the random slope is going to be more appropriate when we have replicates in the data.

Including this random slope also raises an important question, why did we nest the random slopes within subjects and not within stimuli? That is, why "`(1+modality|subject)`" and not "`(1+modality|stimulus)`"? Or why not both random slopes at once? The answer to this question is complicated and we direct readers to Frossard & Renaud (2019) for a more detailed explanation of which random-slopes are allowable. In short, however, we could add a random slope for modality within stimuli in this example because each stimulus was presented in each modality across subjects (although it was presented in only one modality for a given subject). As such, modality varies within stimuli much the same way it varies within subjects, so both random slopes could be included in this example. *However*, this more complicated model fails to converge without some tinkering with different optimizers (see Brown (2021)), so for simplicity we present only the model with "`(1+modality|subject)`".

Broadly speaking, a random slope will be allowable if the slope factor varies within a cluster of observations (i.e., as modality varied within a subjects and within stimuli above). In contrast, a random slope would not be allowable if the slope factor varies between clusters of observations. For example, we could have a random slope of "stimulus length" within subjects because all subjects heard words of various lengths. However, we could not have a random slope of stimulus length within stimuli, because length varies between stimuli not within stimuli.

Table 4. Summary of the model output for the linguistic data, including random intercepts of both participant and stimulus, and a random effect of modality within each participant.

Random Effects				Fixed Effects			
Group	Name	SD	Correlation	Effect	F-Value	df1, df2	p-value
Stimulus	Intercept	0.02		Modality	52.86	1, 51.9	<0.001
Subject	Intercept	0.16					
	Modality	0.07	-0.29				
Residual		0.24					

Note that the number of observations = 21,679, stimuli = 543, and subjects = 53.

Conclusions

First, we would like to thank readers for sticking with us and making it through a rather lengthy monograph. Second, we hope that this gives readers a solid starting point when thinking about what random-effects to include in their mixed-effect model given their study design. Longitudinal designs, factorial repeated measures, and studies with multiple sources of variance all have unique considerations for how to appropriately deal with statistical dependency in the data. We encourage practitioners to work through the vignettes provided in the online repository to get a better practical understanding of how to build these models in R and how to interpret the results. And finally, we hope readers have an appreciation for the complexity one faces when using mixed-effect models. The nuances of these modeling choices are not trivial and can have substantial effects on one's statistical conclusions.

Additional Information

Data Accessibility

Data and code are available via a version controlled online repository: https://doi.org/10.5281/zenodo.10173823

These supplementary materials can also be viewed at: https://keithlohse.github.io/mixed_effects_models/

Author Contributions

- Contributed to conception and design: KRL, AJK, MJS
- Contributed to acquisition of data: NA
- Contributed to analysis and interpretation of data: KRL, AJK, MJS
- Drafted and/or revised the article: KRL, AJK, MJS
- Approved the submitted version for publication: KRL, AJK, MJS

Conflict of Interest

KRL is a member of the educational leadership team for "Reproducible Rehabilitation" (NIH NICHD/NCMRR R25HD105583), teaching content around statistical analysis and mixed-effect models specifically.

Funding

KRL receives salary support as a methodologist from: R01MH123723-01A1, 2 R37 HD068290-10, R01 AR081881, and NICHD R01 HD108240. The authors received no specific financial support for the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to thank many students and trainees who have asked important questions about mixed-effect regression over the years. We would also like to thank two anonymous reviewers whose insightful comments greatly improved the clarity and technical accuracy of the manuscript.

Preprint

The pre-publication version of this manuscript can be found on arxiv, https://arxiv.org/abs/2209.14349

References

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models* . Wiley.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00328

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious Mixed Models* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1506.04967

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). *Fitting Linear Mixed-Effects Models Using {lme4}* (Vol. 67, pp. 1--48). https://doi.org/10.18637/jss.v067.i01

Benjamin, D. J., & Berger, J. O. (2019). Three Recommendations for Improving the Use of p-Values. *The American Statistician*, *73*(sup1), 186–191. https://doi.org/10.1080/00031305.2018.1543135

Bolker, B. (2023). *GLMM FAQ.*

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*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1), 251524592096035. https://doi.org/10.1177/2515245920960351

Brown, V. A., & Strand, J. F. (2019, May 22). *About face: Seeing the talker improves spoken word recognition but increases listening effort.* Center for Open Science. https://doi.org/10.31234/osf.io/m7b8q

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Lawrence Erlbaum Associates Publishers.

Dunlop, D. D. (1994). Regression for Longitudinal Data: A Bridge from Least Squares Regression. *The American Statistician*, *48*(4), 299. https://doi.org/10.2307/2684838

Efron, B., & Tibshirani, R. J. (1993). Bootstrap standard errors: some examples. In *An Introduction to the Bootstrap* (pp. 60–85). Springer US. https://doi.org/10.1007/978-1-4899-4541-9_7

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* CRC Press.

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(2), 399–433. https://doi.org/10.1017/s0080456800012163

Fox, J. (2016). *Applied regression analysis and generalized linear models.* SAGE.

Fox, J., & Weisburg, S. (2011). *An R companion to applied regression.* SAGE.

Frossard, J., & Renaud, O. (2019). *Choosing the correlation structure of mixed effect models for experiments with stimuli* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1903.10766

Garcia, T. P., & Marder, K. (2017). Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model. *Current Neurology and Neuroscience Reports*, *17*(2). https://doi.org/10.1007/s11910-017-0723-4

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, *33*(1). https://doi.org/10.1214/009053604000001048

Goodman, S. N. (2019). Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics. *The American Statistician*, *73*(sup1), 26–30. https://doi.org/10.1080/00031305.2018.1558111

Gurka, M. J., Edwards, L. J., & Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine*, *30*(22), 2696–2707. https://doi.org/10.1002/sim.4293

Hodges, J. S. (2016). *Richly Parameterized Linear Models.* Chapman and Hall/CRC. https://doi.org/10.1201/b16019

Imrie, R. (2004). Demystifying disability: a review of the *International Classification of Functioning, Disability and Health. Sociology of Health & Illness*, *26*(3), 287–305. https://doi.org/10.1111/j.1467-9566.2004.00391.x

Johnson, P. C. D. (2014). Extension of Nakagawa-Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210x.12225

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data Analysis.* Routledge. https://doi.org/10.4324/9781315744131

Kauermann, G., & Carroll, R. J. (2001). A Note on the Efficiency of Sandwich Covariance Matrix Estimation. *Journal of the American Statistical Association*, *96*(456), 1387–1396. https://doi.org/10.1198/016214501753382309

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, *8*(2), 115–128. https://doi.org/10.1037/1082-989x.8.2.115

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics*, *53*(3), 983. https://doi.org/10.2307/2533558

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). *{lmerTest} Package: Tests in Linear Mixed Effects Models* (Vol. 82, pp. 1--26). https://doi.org/10.18637/jss.v082.i13

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*. https://CRAN.R-project.org/package=ez

Lohse, K., Shen, J., & Kozlowski, A. J. (2020, January 29). *Modeling Longitudinal Outcomes: A Contrast of Two Methods*. Center for Open Science. https://doi.org/10.31236/osf.io/b4uev

Long, J. D. (2012). *Longitudinal Data Analysis for the Behavioral Sciences Using R* . SAGE.

Mason, F., Cantoni, E., & Ghisletta, P. (2021). Parametric and semi-parametric bootstrap-based confidence intervals for robust linear mixed models. *Methodology*, *17*(4), 271–295. https://doi.org/10.5964/meth.6607

McLean, R. A., Sanders, W. L., & Stroup, W. W. (1991). A Unified Approach to Mixed Linear Models. *The American Statistician*, *45*(1), 54. https://doi.org/10.2307/2685241

McNeish, D. (2017). Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction. *Multivariate Behavioral Research*, *52*(5), 661–670. https://doi.org/10.1080/00273171.2017.1344538

Nakagawa, S., & Schielzeth, H. (2012). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nelder, J. A. (2007). What is the Mixed-Models Controversy? *International Statistical Review*, *0*(0), 071121035909002–??? https://doi.org/10.1111/j.1751-5823.2007.00022.x

Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer.

Pinheiro, J., Bates, D., & R Core Team. (2022). *nlme: Linear and Nonlinear Mixed Effects Models*. https://CRAN.R-project.org/package=nlme

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209–233. https://doi.org/10.1037/a0020141

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models : applications and data analysis methods*. SAGE.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, *6*(1). https://doi.org/10.1214/ss/1177011926

Sainani, K. (2010). The Importance of Accounting for Correlated Observations. *PMR*, *2*(9), 858–861. https://doi.org/10.1016/j.pmrj.2010.07.482

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*(5), 309–316. https://doi.org/10.1007/bf02288586

Senn, S. (2003). A Conversation with John Nelder. *Statistical Science*, *18*(1). https://doi.org/10.1214/ss/1056397489

Sera, F., Armstrong, B., Blangiardo, M., & Gasparrini, A. (2019). An extended mixed-effects framework for meta-analysis. *Statistics in Medicine*, *38*(29), 5429–5444. https://doi.org/10.1002/sim.8362

Silk, M. J., Harrison, X. A., & Hodgson, D. J. (2020). Perils and pitfalls of mixed-effects regression models in biology. *PeerJ*, *8*, e9522. https://doi.org/10.7717/peerj.9522

Singer, J. D., & Willett, J. B. (2003). Exploring Longitudinal Data on Change. In *Applied Longitudinal Data Analysis* (pp. 16–44). Oxford University PressNew York. https://doi.org/10.1093/acprof:oso/9780195152968.003.0002

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. In *New Methods in Cognitive Psychology* (pp. 4–31). Routledge. https://doi.org/10.4324/9780429318405-2

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. SAGE.

Üstun, T. B., Chatterji, S., Bickenbach, J., Kostanjsek, N., & Schneider, M. (2003). The International Classification of Functioning, Disability and Health: a new tool for understanding disability and health. *Disability and Rehabilitation*, *25*(11-12), 565–571. https://doi.org/10.1080/0963828031000137063

Van der Elst, W., Molenberghs, G., Hilgers, R., Verbeke, G., & Heussen, N. (2016). Estimating the reliability of repeatedly measured endpoints based on linear mixed-effects models. A tutorial. *Pharmaceutical Statistics*, *15*(6), 486–493. https://doi.org/10.1002/pst.1787

Venables, W. N., & Ripley, B. D. (1997). Modern Applied Statistics with S-PLUS. In *Statistics and Computing*. Springer New York. https://doi.org/10.1007/978-1-4757-2719-7

Voss, D. T. (1999). Resolving the Mixed Models Controversy. *The American Statistician*, *53*(4), 352. https://doi.org/10.2307/2686056

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond P $<$ 0.05. *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014