An Introduction to Mixed Models for Experimental Psychology

Henrik Singmann

University of Zurich

David Kellen

Syracuse University

Will appear as:

Singmann, H., & Kellen, D. (in press). An Introduction to Mixed Models for Experimental Psychology. In D. H. Spieler & E. Schumacher (Eds.), *New Methods in Neuroscience and Cognitive Psychology*. Psychology Press.

Author Note

We thank Ben Bolker, Jake Westfall, and Rene Schlegelmilch for very helpful comments on a previous versions of this chapter. Henrik Singmann and David Kellen received support from the Swiss National Science Foundation Grant 100014_165591.

In order to increase statistical power and precision, many psychological experiments collect more than one data point from each participant, often across different experimental conditions. Such *repeated-measures* pose a problem to most standard statistical procedures such as ordinary least-squares regression or (between-subjects) ANOVA as those procedures assume that the data points are *independent and identically distributed* (henceforth *iid*). The iid assumption is comprised of two parts: The assumption of identical distribution simply means that all observations are samples from the same underlying distribution. The independence assumption states that the probability of a data point taking on a specific value is independent of the values taken by all other data points.[1] In this chapter we are mainly concerned with the latter assumption.

It is easy to see that in the case of repeated measures the independence assumption is expected to be violated. Observations coming from the same participant are usually correlated; e.g., they are more likely to be similar to each other than two observations coming from two different participants. For example, when measuring response latencies a participant that is generally slower than her/his peers will respond comparatively slower across conditions, thus making the data points from this participant correlated and non-independent (i.e., a participant's rank in one condition is predictive of their rank in other conditions). More generally, one can expect violations of the iid assumption if data are collected from units of observations that are clustered in groups. Other examples of this are data from experiments collected in group settings, students within classrooms, or patients within hospitals. In such situations one would expect that observations within each cluster (i.e., a specific group, classroom, or hospital) are more similar to each other than observations across clusters.

[1] Technically, the independence assumption does not pertain to the actual data points (or marginal distribution), but to the residuals (or conditional distribution) once the statistical model (i.e., fixed effects, random effects, etc.) has been taken into account. With this, we can define independence formally via conditional probabilities. The probability that any observation i takes on a specific value x_i is the same irrespective of the values taken on by all the other observations $j \neq i$, and a statistical model with parameter vector $\boldsymbol{\theta}$: $P(x_i|\boldsymbol{\theta}) = P(x_i \mid \boldsymbol{\theta}, \bigcap_{j \neq i} x_j)$.

Unfortunately, compared to violations of other assumptions, such as the normality assumption or the assumption of variance heterogeneity in ANOVA, standard statistical procedures are usually not robust to violations of the independence assumption (Judd, Westfall, & Kenny, 2012; Kenny & Judd, 1986). In a frequentist statistical framework such violations often lead to considerably increased Type I errors (i.e., false positives). More generally, such violations can produce overconfident results (e.g., too narrow standard errors).

In this chapter we will describe a class of statistical model that is able to account for most of the cases of non-independence that are typically encountered in psychological experiments, *linear mixed effects models* (LMM, e.g., Baayen, Davidson, & Bates, 2008), or mixed models for short. Mixed models are a generalization of ordinary regression that explicitly capture the dependency among data points via random effects parameters. Compared to traditional analyses that ignore these dependencies, mixed models provide more accurate (and generalizable) estimates of the effects, improved statistical power, and non-inflated Type I errors. The reason for the recent popularity of linear mixed models boils down to the computational resources required to implement them: In the absence of such resources, realistic data-analytic methods had to rely on simpler models that ignored the dependencies in the data, and relied on closed-form estimates and asymptotic results. Fortunately, today we can easily implement most linear mixed models using any recent computer with sufficient RAM.

The remainder of this chapter is structured as follows: First, we introduce the concepts underlying mixed models and how they allow to account for different types of non-independence that can occur in psychological data. Next, we discuss how to set up a mixed model and how to perform statistical inference with a mixed model. Then, we will discuss how to estimate a mixed model using the `lme4` (Bates, Mächler, Bolker, & Walker, 2015) as well as the `afex` (Singmann, Bolker, Westfall, & Aust, 2017) packages for the statistical programming language `R` (R Core Team, 2016). Finally, we will provide an outlook of how to extend mixed models to handle non-normal data (e.g., categorical responses).

Fixed Effects, Random Effects, and Non-Independence

The most important concept for understanding how to estimate and how to interpret mixed models is the distinction between *fixed* and *random effects*.[2] In experimental settings fixed effects are often of primary interest to the researcher and represent the overall or *population-level average* effect of a specific model term (i.e., main effect or interaction) or parameter on the dependent variable, irrespective of the random or stochastic variability that is present in the data. A statistically-significant fixed effect should be interpreted in essentially the same way as a statistically-significant test result for any given term in a standard ANOVA or regression model. Furthermore, for fixed effects one can easily test specific hypotheses among the factor levels (e.g., planned contrasts).

In contrast, random effects capture random or stochastic variability in the data that comes from different sources, such as participants or items. These sources of stochastic variability are the *grouping variables* or *grouping factors* for the random effects and always concern categorical variables (i.e., nominal variables such as condition, participant, item) – continuous variables cannot serve as grouping factors for random effects. In experimental settings, it is often useful to think about the random effects grouping factors as the part of the design a researcher wants to generalize over. For example, one is usually not interested in knowing whether or not two factor levels differ for a specific sample of participants (after all, this could be done simply by looking at the obtained means in a descriptive manner), but whether the data provides evidence that a difference holds in the population of participants the sample is drawn from. By specifying random effects in our model, we are able to factor out the idiosyncrasies of our sample and obtain a more general estimate of the fixed effects of interest.[3]

[2]Note that there are different possibilities on how to define fixed and random effects, ways that are not necessarily compatible with each other (Bolker, 2015; Gelman, 2005). The definition employed here is the one most useful for understanding how to specify and estimate frequentist mixed model as implemented in `lme4` (Bates, Mächler, et al., 2015).

[3]It should be noted that this distinction, fixed effects as variables of interests versus random effects as nuisance variables one wants to generalize over, is a simplification. Fixed effects can also serve as

The independence assumption of standard statistical models implies that one can only generalize across *exactly one* source of stochastic variability: the population from which each observation (i.e., row in most statistical software packages) is sampled. In psychology the unit of observation is usually participants, but occasionally other units such as items are employed alternatively (e.g., having words as the unit of observation is fairly common in psycholinguistic research). Importantly, the notion that the unit of observation represents a random effect is usually only an implicit part of a statistical model. In contrast, mixed models require an explicit specification of the random-effects structure embedded in the experimental design. As described above, the benefit of this extra step is that one can adequately capture a variety of dependencies that standard models cannot.

In order to make the distinction and the role of random effects in mixed models clearer, let us consider a simple example (constructed after Baayen et al., 2008 and Barr, Levy, Scheepers, & Tily, 2013). Assume you have obtained response latency data from I participants in $K = 2$ difficulty conditions, an *easy condition* that leads to fast responses and a *hard condition* that produces slow responses. For example, in both conditions participants have to make binary judgments on the same groups of words: In the easy condition they have to make animacy judgments (whether it is a living thing). In the hard condition participants have to a) judge whether the object the word refers to is larger than a soccer ball and b) whether it appears in the northern hemisphere; participants should only press a specific key if both judgments are positive. Moreover, assume that each participant provides responses to the same J words in each difficulty condition. Thus, difficulty is a repeated-measures factor, more specifically a within-subjects factor with J replicates for each participant in each cell of the design, but also a within-words factor with I replicates for each word in each cell of the design. Note that the cells of a designs are given by the combination of all (fixed) factor levels. In the present example there are two cells, corresponding to the easy condition and the

---

nuisance variables (e.g., when including a fixed effect to "statistically control" for it) and random effects (e.g., intra-class correlations) can be of primary interest.

difficult condition, but in a 2 × 2 design we would have four cells instead.[4]

Figure 1 illustrates the response latency data from two participants (S_1 and S_2) across the easy and hard conditions, for two words (I_1 and I_2). The different panels show the observed data together with the predictions from a specific model. Going from top to bottom, the complexity of these models increases. The features of each of these models will become clear throughout the remainder of this chapter. But at this point a brief description of the data is in order: First, note that there is a general individual difference across conditions, with Subject 1 being overall slower than Subject 2. Also, the two subjects differ in terms of the slowing-down effect observed between the easy and hard conditions, with the increase in response latency being larger for Subject 2 than for Subject 1. We also find that responses to item I_1 tend to be generally faster than I_2, a difference that is smaller in the hard condition. The models discussed below will differ in their ability to account for these differences observed across subjects and items.

Fixed-Effects-Only Model

Let us first consider the ordinary regression model that completely ignores the non-independence in the data. Such a model could be specified in the following manner:

$$\begin{aligned} y_{i,j,k} &= \beta_0 + \beta_\delta X_{i,j,k} + \epsilon_{i,j,k}, \\ & i = 1, 2, \ldots, I, \qquad j = 1, 2, \ldots, J, \qquad k = 1, 2, \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma_\epsilon^2), \end{aligned} \tag{1}$$

where $y_{i,j,k}$ denotes the dependent variable (here observed response latencies) for the ith participant and jth item in the kth condition. Parameter β_0 is the intercept and grand mean, β_δ corresponds to the effect of the difficulty condition, and $X_{i,j,k}$ is an indicator variable which takes on value 1 in the easy condition and -1 in the hard condition (thus

[4]For simplicity the example assumes perfect *balance* (i.e., all cells have the same number of participants, items, and observations). In principle, the methods discussed here generalize to non-balanced data sets, but numerical or other issues often arise if the imbalance is not small. Furthermore, imbalance can considerably impact power (Judd, Westfall, & Kenny, 2017).

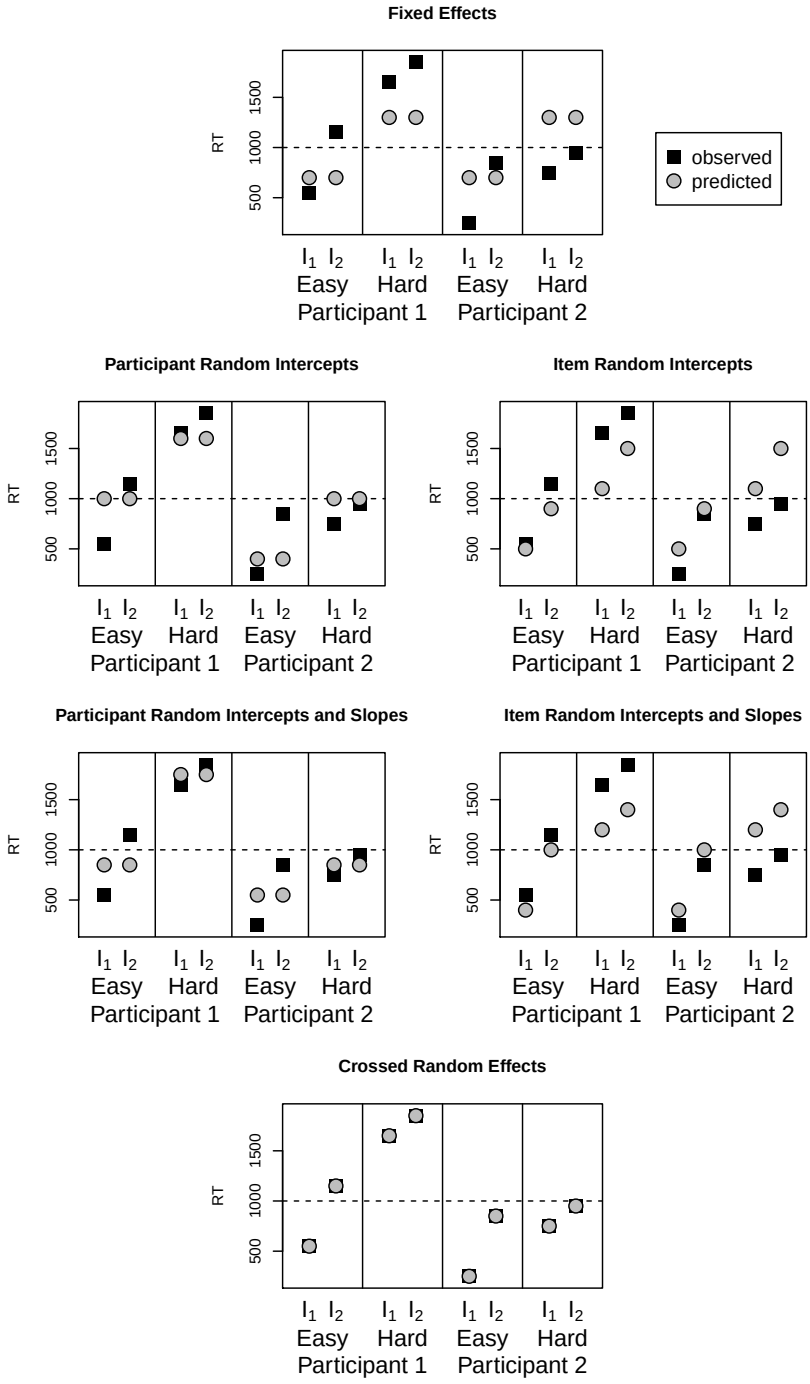


Figure 1. Example data and the predictions from different models. The complexity of the models increases across rows (top row being the simplest model). Note that this is an illustration only; a model that perfectly describes the data as shown in the bottom row is non-identifiable as it has more parameters than data points.

$2 \times \beta_\delta$ is the difference between the conditions). Finally, $\epsilon_{i,j,k}$ is the residual error. The third row states that the vector of all residual errors, $\boldsymbol{\epsilon}$ (non-scalar variables such as vectors or matrices are printed in **bold** font), is assumed to follow a normal (i.e., Gaussian: $\mathcal{N}$) distribution with mean 0 (i.e., zero-centered) and residual variance σ^2_ϵ. This distribution of residuals implies the iid assumption, which is clearly violated for the given data. For example, the fact that Subject 1's response in the easy condition for item I_1 is slower than Subject's 2 response to the same item is predictive of their relative speed in the hard condition. Overall, the fixed-effects model provides a poor account of the data as it completely precludes any of the dependencies that are present in it (see Figure 1, top row).

Random-Intercepts Model

As noted above, independence is violated if we can learn something about a specific data point by knowing the value of a different data point (after taking the structure of the statistical model into account). A natural assumption here would be that data points from one participant are more similar to each other than data points from other participants. One way to interpret this assumption is to assume that each participant has an idiosyncratic overall response latency; some participants are slower than the average and some are faster than the average. At this point, the shortcomings of the statistical model described in Equation 1 becomes clear: it only assumes a single intercept β_0 to characterize all participants.

In order to allow for idiosyncratic average response latencies per participant we need to introduce effects that capture the displacement of each participant from the grand mean (i.e., the intercept β_0). Such a model could be specified as:

$$\begin{aligned} y_{i,j,k} &= \beta_0 + S_{0,i} + \beta_\delta X_{i,j,k} + \epsilon_{i,j,k}, \\ & i = 1, 2, \ldots, I, \qquad j = 1, 2, \ldots, J, \qquad k = 1, 2, \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma^2_\epsilon), \\ \boldsymbol{S_0} &\sim \mathcal{N}(0, \sigma^2_{S_0}), \end{aligned} \tag{2}$$

where $S_{0,i}$ corresponds to the idiosyncratic effect associated to participant i.

Furthermore, we assume that the vector of the idiosyncratic effects, $\boldsymbol{S_0}$, follows a zero-centered normal distribution with variance $\sigma^2_{S_0}$. The individual $S_{0,i}$ values can be either positive and negative, summing up to zero. These values allow individual participants to have their own "grand means" (see Figure 1, second row), which are assumed to be normally distributed around β_0.

Because this model adjusts the grand mean or intercept β_0 for each participant, this model is commonly referred to as a *random-intercepts model*. In this particular case the traditional fixed-effects model is augmented with *by-participant random intercepts* (i.e, participant is the random effects grouping factor for which we estimate random intercepts). Note that the random-intercepts model is sufficient to account for correlations across data points that are brought about by differences in overall levels of performance such as some participants being generally slower and some being generally faster.

From the model expressed in Equation 2 it is relatively easy to see the differences between the fixed effects and the random effects. The fixed effects parameters are the traditional regression parameters β_0 and β_δ. Both of these are scalar values; there is exactly one value for β_0 which represents the grand mean across all participants (for our example the total mean response time) and one value for β_δ which represents the difference between the two condition across all participants. In contrast, the random effects vector $\boldsymbol{S_0}$ includes all the idiosyncratic displacements $S_{0,i}$ that are added to the grand mean for each participant. It is is important to keep in mind that this model only has four parameters: β_0, β_δ, σ^2_ϵ, and, $\sigma^2_{S_0}$. This number does not change as a function of the number of individuals (e.g., more parameters as sample size increases). In order to specify the random intercepts, it suffices to introduce the variance parameter $\sigma^2_{S_0}$. The participant-level displacements captured by $\boldsymbol{S_0}$ do not correspond to parameters that were estimated (e.g., unlike β_0). Instead, they correspond to *conditional modes* (sometimes referred to as posterior mean values) obtained when conditionalizing on the data and the above-described parameters (see Bates, Mächler, et al., 2015, Pinheiro & Bates, 2000; the individual displacements are also known as *best linear unbiased*

predictions or *BLUPs*, but this terminology is somewhat outdated).

Random-Intercepts-and-Slopes Model

The random-intercepts model expressed in Equation 2 does not, however, account for all potential dependencies in the data brought about by the different participants. The reason for this is the presence of the within-subjects factor 'difficulty'. The two previous models assume that the difference between the difficulty conditions is equal for all participants, but that does not necessarily have to be the case. It is easy to imagine that this difference is larger for some participants but smaller for other participants (or even takes on a different direction). For example, imagine that for some participants the conjunctive task associated to the hard condition is particularly challenging, leading to larger differences between the two conditions. This situation would lead to dependencies that would be unaccounted by the models discussed so far: by knowing the values in a pair of data points from such a participant (one from each condition), we know something about other possible pairs of data points.

In order to account for such dependencies at the level of a given factor, we once again introduce a new random effect corresponding to the participant-level displacements from the condition effect β_δ:

$$\begin{aligned} y_{i,j,k} &= \beta_0 + S_{0,i} + (\beta_\delta + S_{\delta,i})X_{i,j,k} + \epsilon_{i,j,k}, \\ & i = 1, 2, \ldots, I, \qquad j = 1, 2, \ldots, J, \qquad k = 1, 2, \\ & \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2), \\ \begin{pmatrix} \boldsymbol{S_0} \\ \boldsymbol{S_\delta} \end{pmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{S_0}^2 & \rho_{S_0,S_\delta}\sigma_{S_0}\sigma_{S_\delta} \\ \rho_{S_\delta,S_0}\sigma_{S_0}\sigma_{S_\delta} & \sigma_{S_\delta}^2 \end{bmatrix} \right), \end{aligned} \tag{3}$$

where $S_{0,i}$ is the displacement of participant i from β_0, and $S_{\delta,i}$ is displacement of the same participant i from the difficulty effect β_δ (see Figure 1, third row). We now estimate two random-effect vectors, the random intercept $\boldsymbol{S_0}$ and a random effects term added to the condition effect, $\boldsymbol{S_\delta}$. Furthermore, we assume that the two random effects come from a zero-centered multivariate normal distribution for which we estimate both variances, $\sigma_{S_0}^2$ and $\sigma_{S_\delta}^2$, as well as the correlation, $\rho_{S_\delta,S_0} = \rho_{S_0,S_\delta}$. Estimating the

correlation allows us to account for dependencies that arise when both random effects are correlated. For example, participants that are overall slow could also have an overall larger condition effect which would result in a positive correlation. Because the regression parameters besides the intercept are usually called slopes, $\boldsymbol{S_\delta}$ is also known as a *random slope*. Thus, the model in Equation 3 is a mixed model with by-participant random-intercept and by-participant random slopes as well as a correlation among the by-participant random effects.

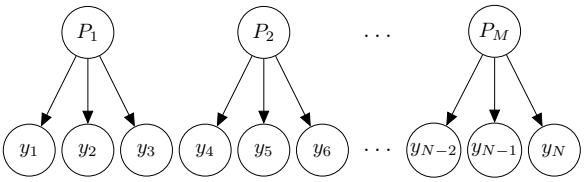
We have now established the fundamental distinction between fixed and random effects in mixed models, and discussed the different types of random effects that can be added to a model: random intercepts, random slopes, and correlations among random effects for a specific random effects grouping factor. The random effects are added to a model so that it can provide a more accurate account of the data-generating process that takes into account the heterogeneity observed across participants as well as the dependencies that are expected in the data. The shortcomings of the random-intercepts model expressed in Equation 2 and the extension expressed in Equation 3 clarifies the need to include a random slope per factor in order to account for the possibility that the differences observed across the levels of a factor can vary across participants. Failure to add such random slopes can lead to considerably increased Type I error rates as discussed in greater detail below (Barr et al., 2013; Schielzeth & Forstmeier, 2009).

It is important to keep in mind that *random effects do not alter the interpretation of the fixed effects*. If we are interested in knowing whether a specific factor has an overall effect, this is only possible by investigating the fixed effects. The random effects only tell us whether or not there is variation in a fixed effect for the different levels of the random effects term (from this it follows that it is rare to include a random effects parameter, but not the corresponding fixed effect). But given that the variation is zero-centered, the random effects cannot adjust the overall effect. They are added with the sole purpose of accounting for the non-independence present in the data due to observing multiple observations from a given random-effects level. Also important to note is the fact that the introduction of random effects does not necessarily translate

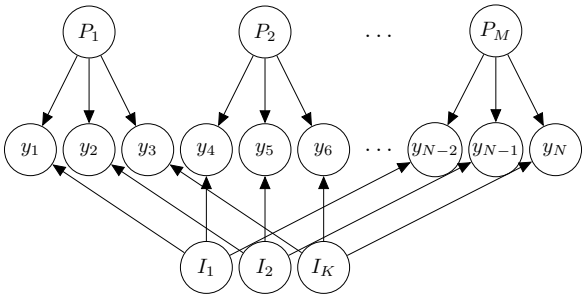
into a considerable increase in the number of parameters to be estimated (one variance parameter per effect). However, the same does not hold when the correlations across effects are also estimated. The model in Equation 3 including random intercepts and random slopes only introduces one correlation to be estimated. But as the number of random effects increases, the number of correlations to be estimated can explode. For a model with r random effects, $r(r-1)/2$ correlations could be estimated (e.g., for $r = 5$, the number of correlations is 10).

There is one more angle from which to view random effects, namely how they allow us to improve estimation on an individual level. The simplest model to for the example data was the simple regression model given in Equation 1. If one ignores all individual variability and dependencies and estimates the complete data set with this model one employs *complete pooling*, with all data points being treated as independent observations, which violates the iid assumption. An alternative analysis strategy that takes the dependencies and individual variability into account would be to fit the data of each individual participant separately with the model in Equation 1. With this approach, one would obtain an individual set of regression parameters for each participant, which could then be analyzed in a second step. This approach, designated as *no pooling*, would not violate the iid assumption. However, this no-pooling approach has a few downsides: a) It does not easily lend itself to mixed designs with both between- and within-subjects factors, b) one has to decide on how to analyze the individual parameter estimates, and c) it requires sufficient data on the individual level for obtaining reliable parameter estimates. Mixed models, with both fixed and random effects, provide a principled compromise via what is known as *partial pooling*: The random effects allow each individual participants to basically have an individual parameter estimate, as in the no pooling approach. However, the normality assumption underlying the random effects provides additional structure which ensures that the estimation of each individual participants' data is informed by the complete data set. Mixed models therefore acknowledge the presence of individual differences, but at the same time take into account the fact that individuals are similar to a certain degree.

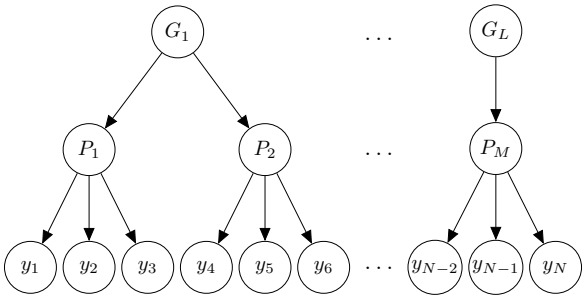
Figure 2. Different Types of Random Effects. Observations are labeled from y_1 to y_N. Different participants are labeled from P_1 to P_M. Different items are labeled from I_1 to I_K. Different groups (where each participant is in exactly one group) are labeled from G_1 to G_L.



(a) Single Random Effect



(b) Crossed Random Effects



(c) Nested Random Effects

The normality assumption also leads to what is known as *hierarchical shrinkage*: the individual parameter estimates for individuals for which the data diverges considerably from the rest are adjusted towards the mean. As a consequence – and in contrast with traditional ANOVA models – the predicted cell means of mixed models do not necessarily coincide with the observed cell means.

Crossed and Nested Random Effects

One important characteristic of mixed models is that they allow random effects for multiple, possibly independent, random effects grouping factors. Figure 2 provides an overview over the different random effects grouping factor types discussed in this chapter. In the models expressed in Equations 2 and 3 we only introduced by-participant random effects (Figure 2a). However, participants were not the only source of stochastic variability in the example experiment. The task was to judge words and each presented word can also be seen as a sample from the population of all words.

Therefore, the sample of presented words can be seen as another source of stochastic variability (Clark, 1973). Figure 1 (second and third row) illustrates how the fixed-effects model could have been extended with random intercepts and slopes for items rather than for subjects. However, in experiments in which both participants and items are sampled, there is interest in simultaneously generalizing across the two sources of stochastic variability and not only one of them.

Generalization across both participants and items can be easily achieved by adding by-item random effects in addition to the by-participant random effects. Note that in the example experiments, condition varies within words (i.e., each word appears in each difficulty condition) and we thus not only want to have by-item random intercepts but also by-item random slopes (Figure 2b). The full mixed model is given by:

$$
\begin{aligned}
y_{i,j,k} &= \beta_0 + S_{0,i} + I_{0,j} + (\beta_\delta + S_{\delta,i} + I_{\delta,j}) X_{i,j,k} + \epsilon_{i,j,k}, \\
& i = 1, 2, \ldots, I, \qquad j = 1, 2, \ldots, J, \qquad k = 1, 2, \\
\boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma^2_\epsilon), \\
\begin{pmatrix} \boldsymbol{S_0} \\ \boldsymbol{S_\delta} \end{pmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{S_0} & \rho_{S_0,S_\delta}\sigma_{S_0}\sigma_{S_\delta} \\ \rho_{S_\delta,S_0}\sigma_{S_0}\sigma_{S_\delta} & \sigma^2_{S_\delta} \end{bmatrix} \right), \\
\begin{pmatrix} \boldsymbol{I_0} \\ \boldsymbol{I_d} \end{pmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{I_0} & \rho_{I_0,I_\delta}\sigma_{I_0}\sigma_{I_\delta} \\ \rho_{I_\delta,I_0}\sigma_{I_0}\sigma_{I_\delta} & \sigma^2_{I_\delta} \end{bmatrix} \right),
\end{aligned}
\tag{4}
$$

where $\boldsymbol{I_0}$ is the by-item random intercept and $\boldsymbol{I_\delta}$ the by-item random slope for the difficulty effect for the J different words. As before, for each by-item random effect we estimate the corresponding variance, here $\sigma^2_{I_0}$ and $\sigma^2_{I_\delta}$, as well as their correlation ρ_{I_0,I_δ}. Because each item appears for each participant, the corresponding random effects are known as *crossed* (this would also be the case if each participant worked on a random subset of the items). In the example shown in Figure 1, the crossed random effects model including random intercepts and slopes for participants and items is able to perfectly capture the data.[5]

[5]Please note that more data points than shown in Figure 1 are necessary for uniquely identifying the parameters of the model and a perfect description of the data is usually not possible for a model with identifiable parameters, see below for details.

The mixed model in Equation 4 with crossed random effects for participants and items and by-participant and by-item random intercepts as well as random slopes for the difficulty effect as well as correlation among the by-item random effects and correlation among the by-participant random effects implements the *maximal random effects structure justified by the design*.[6] As already alluded to earlier, by virtue of being able to accommodate different forms of heterogeneity that can be found in the data, the maximal model is the mixed model that is the most likely to provide the accurate description of the data generating process in the mixed model framework and the model that in principle provides the best protection against inflated Type I errors (Barr et al., 2013; Schielzeth & Forstmeier, 2009). It is therefore always a good idea to start a mixed model analysis of a new data set with the maximal model. We will return to the question what to do should this model not converge successfully below.[7]

One common problem in designs with crossed-random effects is the identification of the maximal random effects structure. For the by-participant random effects grouping factor the maximal structure is simply the combination of all within-subjects factors (i.e., main effects and all interactions). For identifying the maximal structure for a (crossed) by-item random effects grouping factor it is important to temporarily *ignore* the different participants. The maximal structure of the random item effect is then simply the combination of all factors varying within items. More specifically, if a

[6]If we had replicates for each combination of participant and item, we could additionally estimate random effects for the random effects grouping factor resulting from the participant-by-item interaction, potentially with random slopes for all fixed effects. With such data, such a model would constitute the maximal model. As the example experiment however did not include this data, this effect is confounded with the residual variance and cannot be estimated.

[7]It should be noted that the modeling of crossed-random effects supersedes remedial approaches that are commonly used by researchers. Prominent among these, particularly in the psycholinguistic literature, is the separate testing of effects when aggregating data across participants versus aggregating across items (Clark, 1973). Instead of checking the robustness of an effect when relying on alternative aggregation procedures (e.g., F_1 vs. F_2 tests), linear mixed models do not rely on aggregation at all and estimate the different dependencies in the data directly. This not only avoids ambiguities on how to best integrate the different analysis, but also provides a better protection against inflated Type I errors (Barr et al., 2013).

specific item appears in all levels of a given factor (independent of whether that happens within one participant or across participants) this factor varies within items and the corresponding random slope is part of the maximal random effects structure. And if two different factors are both within-item factors, the random slope of the interaction of the factors is also part of the maximal random effects structure. Note further that for unbalanced designs (e.g., if the items for each participant are selected randomly from a pool of items) to be able to estimate by-item random slopes for a specific fixed fixed effect it is not necessary that this fixed effect varies among all levels of the item effect. In other words, mixed models can deal with missing values on the level of the random effects. If there is too much missing data the variance of the random effects will simply be estimated to be zero.

Some random effects are not crossed, but *nested*; this is the case if some levels of one random effects grouping factor only appear in one level of a second random effects grouping factor (Figure 2c). For example, if participants are tested in groups, participants are nested within groups, and if different groups are tested by different experimenters, groups itself would again be nested within experimenters. In such designs, the lower level grouping factors (i.e., participants) are sometimes called "Level 1", the second lowest (i.e., groups) "Level 2", etc., and statistical models with nested random effects are also called multilevel or hierarchical models. It is important to realize that in terms of the statistical model, crossed and nested random effects do not behave differently. One can only add random intercepts, random slopes, and their correlations for any random effects grouping factor. For example, if in our study participants were collected in groups, we could add a by-group random effects such as by-group random intercepts, by-group random slopes, and their correlation to the model presented in Equation 4 to account for potential correlations among participants that were tested in the same group. In contrast to experimental designs with crossed random effects in which the random effects can often be considered nuisance parameters and are not of primary interest, researchers are often interested in the values of the random effects parameters in nested designs. More specifically, researchers are often

interested in the *intraclass correlation coefficient* (ICC) which is a measure of the degree of similarity within the levels of the random effect. A comprehensive overview about the specific type of statistical and substantive questions that are of interest in nested designs is provided by Snijders and Bosker (2012).[8]

You should now have a clear understanding of the difference between fixed and random effects, but let us sum this up once again. Random effects (or random effects parameters) are zero centered offsets or displacements that are added to the fixed effect parameters in order to control for dependencies in the data, ultimately providing a more accurate description of the data-generating process. An important rule of thumb is that a random effects can only be specified practically for grouping factors which have at least five or six different levels (Bolker, 2015). With fewer levels the estimation of the variance of the random effects will be very imprecise which can lead to numerical problems in the estimation process. Random effects with a low number of levels also have an extremely detrimental effect on statistical power (Westfall, Kenny, & Judd, 2014). This goes so far that for crossed random effects the random effects grouping factor with the lower number of levels provides an upper bound of the maximally attainable power. For example, in an experiment with crossed random effects for participants and items but only 8 different items the maximally attainable power is below .5 even under otherwise favorable conditions and with unlimited participants (Judd et al., 2017, Figure 2). Westfall et al. (2014) provide two rules of thumb for power in mixed models (pp. 2033): (1) "it is generally better to increase the sample size of whichever random factor is contributing more random variation to the data" and (2) "if one of the two sample sizes is considerably smaller than the other, there is generally a greater power benefit in increasing the smaller sample size compared to the larger

[8]Note that not all computer programs fully support all types of random effects. Specifically, older programs or specialized and/or niche programs sometimes only support nested random effects and do not support (or at least not fully) crossed random effects (e.g., in `R` this is true for package `nlme`). For the software discussed here there is usually no necessity to treat crossed and nested random effects differently as long as the levels of each random effects grouping factor receive unique identifiers (e.g., the identifier "participant 1" only exists exactly once and not in two different groups).

sample size". In line with the recommendation of Simmons, Nelson, and Simonsohn (2011) that each between-subjects condition should have at least 20 participants we recommend that each random effects grouping factor should have at least 20 levels, otherwise the power is likely too low.

Another important aspect is that one can only estimate a specific random effects parameter if there are multiple observations for each level of the random effects grouping factor and the fixed effects parameter to which one wants to add the random effects parameter. If this is not the case (i.e., there is only one observation for each level of the random effects grouping factor), the random effects parameter is confounded with the residual variance and cannot be uniquely identified (see also Footnote 6). For example, if each participant provides only one observation in total (i.e., a completely between-subjects design) one cannot even estimate by-participant random intercepts and consequently no mixed model. Likewise, if one only has one observation of each participant in each within-subject condition (as is the case in a traditional repeated-measures ANOVA), one cannot estimate by-participant random slopes for that condition. Mixed models require replicates (i.e., multiple observations) for each level of the random effects grouping factor and each factor that varies within the random effect.

Setting up a Mixed Model

Before discussing the software implementations for fitting mixed models, we still need to discuss a few issues: How to perform statistical inference, how to set up the random effects structure, how to deal with categorical independent variables (i.e., factors), and effect sizes.

Statistical Inference in Mixed Models

Statistical inference (i.e., obtaining *p*-values concerning null hypotheses) in the context of mixed models is far from being a trivial endeavor. The main problem is, again, the existence of dependencies within levels of a random effects grouping factor. This prevents a simple counting of the denominator degrees of freedom via the number of observed data points, as done in standard ANOVA. As a consequence, the standard `R`

function for mixed models, `lmer`, does not report any p-values (Bates, 2006). However, we now have several methods that allow for us to obtain p-values (for an overview see https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#testing-hypotheses). Here, we mainly focus on those methods that compare two nested models. Nested models means that one of the two models, the reduced model, needs to be a special case of the other model, the encompassing model (i.e., the reduced model corresponds to the encompassing model when one or several parameters from the latter are 0). Specifically, in order to test a specific effect (i.e., main effect or interaction) the encompassing model is usually the full model that includes all parameters and the reduced model is the one in which the parameters corresponding to that effect are withheld (i.e., fixed to 0). More specifically, the fixed effects parameters and not the random effects parameters are withheld. We generally recommend the *Kenward-Roger approximation* (Halekoh & Højsgaard, 2014; Kenward & Roger, 1997), which is based on a modified F test and also estimates the denominator degrees of freedom, as it is known to provide the best control of Type I errors with the limited sample sizes that are common in experimental designs in psychology. However, the Kenward-Roger approximation is the most expensive method in terms of computational resources. Especially with complicated random-effect structures (i.e., many random slopes and correlations among random parameters) it may require amounts of RAM that can exceed what is available in normal computers. An alternative that is less expensive in terms of RAM, but quite similar in terms of Type I error control, is the Satterthwaite approximation (Kuznetsova, Brockhoff, & Christensen, 2016; Satterthwaite, 1941).

An alternative that does not rely on approximating the denominator degrees of freedom is the *likelihood ratio test* (LRT). The LRT is a standard statistical test for comparing the goodness of fit of two nested models. This test consists of the ratio of the maximum likelihoods of the encompassing and reduced models.[9] The test statistic of the LRT follows asymptotically the χ^2-distribution with degrees of freedom equal to the

[9]Note that the Kenward-Roger approximation requires a model to be fitted with restricted maximum-likelihood estimation (REML), for details see (Bates & DebRoy, 2004).

difference in number of parameters between encompassing and reduced model. In other words, only in the limit with unlimited levels for the random effects grouping factors does the LRT adequately control for Type I errors. With limited sample sizes the LRT tends to be anti-conservative (i.e., gives significant p-values although the null hypothesis is true; e.g., Pinheiro & Bates, 2000) and we can only recommend its use if the number of levels for each random effects grouping factor is considerable (e.g., $>$ 40 or 50).

In case one does not want to rely on the asymptotic nature of the LRT, a further alternative is *parametric bootstrap*. The parametric bootstrap procedure simulates synthetic datasets from the reduced model and then fits both models to the synthetic data which produces a reference distribution of likelihood-ratio values under the null-hypothesis. The parametric bootstrap p-value corresponds to the percentage of simulated likelihood-ratio values that are larger than the observed likelihood-ratio value. One potential problem with the parametric bootstrap is that for complex models, calculating the sampling distribution of the likelihood-ratio test under the null hypothesis can be quite time consuming. However, it should be able to control for Type I error better than the LRT. Note that parametric bootstrap and the LRT are procedures that can also be used to test the parameters associated to the random effects (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017; see also Scheipl, Greven, & Küchenhoff, 2008).

Another alternative is to simply compare the t-statistic of the parameter estimates with the z distribution (e.g., Baayen, 2008; the rule of thumb is that values larger than 2 indicate statistical significance). Unfortunately, this approach has two problems. First, it can only be used with factors with two levels. As soon as a factor has more than two levels, inspecting the parameter estimates becomes very challenging and essentially useless if the factor is part of an interaction. Second, this approach does the worst job in controlling for Type I errors. We therefore cannot generally recommend to perform statistical inference for mixed models using this "t as z" approach.

Specifying the Random Effects Structure

Perhaps more important than the choice of method for evaluating statistical significance is the correct specification of the random-effects structure. Omitting a random effect when there is in fact variability in this effect across the levels of a random effects grouping factor can dramatically increase Type I errors as shown in a number of independent simulation studies (Barr et al., 2013; Judd et al., 2012; Schielzeth & Forstmeier, 2009). This means that in most cases one should initially start with the *maximal random effects structure* justified by the design as recommended by Barr et al. (2013). The maximal model is the model that includes random effects parameters for all sources of stochastic variability (i.e., random effects grouping factors). Specifically, it contains random intercepts as well as random slopes for all fixed effects that vary within the levels of a given random effects grouping factor, plus the correlations among the random effects.

For the limited sample sizes that are common in psychology and related disciplines a common problem is that the maximal model is not fully identified (Bates, Kliegl, Vasishth, & Baayen, 2015), especially for mixed models with complicated random effects structures. Even though the optimization algorithm converges to the optimum (i.e., the maximum-likelihood estimate for a given data set) the variance-covariance matrix of the random effects parameters at the optimum is degenerate or singular. At least for models estimated with `lme4` this is often signified by convergence warnings. Other signs of singular fits are variance estimates of or near zero and correlation estimates of ± 1. The occurrence of such situations is due to the fact the parameters associated to random effects (e.g., $\sigma^2_{S_\delta}$) are more difficult to estimate than fixed effects (e.g., β_δ). Additional ways to detect degenerate fits are discussed in Bates, Kliegl, et al. (2015).

In the case of a singular fit, it is in principle recommended to reduce the random effects structure given that degenerate or overparameterized models can reduce the statistical power of any tests conducted with them (Matuschek et al., 2017). As a first step, it seems advisable to remove the correlations among random slopes as these contribute the largest number of random effects parameters if the number of variance

parameters for a given random effects grouping factor exceeds three. Additionally, the correlation parameters appear to be even more difficult to estimate than the variance parameters, at least in a frequentist framework (Bates, Kliegl, et al., 2015). If a model still shows problems after removing correlations, other random-effects parameters could be removed, starting with the highest-order random effects parameter with the lowest estimated variance. Empirical-based approaches to address this question and obtain the *optimal* random effects are provided by Matuschek et al. (2017) and Bates, Kliegl, et al. (2015). However, following these recommendation usually requires the researcher to choose a specific model selection procedure and criterion. The consequence of this is that researchers that decide to report results based on a reduced model should make this explicit to the reader and be prepared to defend their choices. In any case, we recommend that one should always start with the maximal model and reduce random effects instead of starting with a minimal model and gradually include random effects.

One problem that may arise from an iterative procedure for determining the random effects structure is that sometimes it might not be possible to reduce the random-effects structure such that all problematic random effects parameters can be removed (e.g., in cases when there is random variability in higher-order effects, but not in lower-order effects). From the pragmatic standpoint that false positives are in principle more dangerous for scientific progress than false negatives, we suggest that in those cases one can accept a few problematic or degenerate parameters (e.g., variances of zero). This strategy strikes us as more reasonable (i.e., more conservative) than simply removing justifiable random effects and inflating Type I error rates to an unknown degree. It is clear that a model with such problematic or degenerate parameters is not the most adequate from a purely statistical standpoint, but it can nevertheless be a reasonable solution if the focus is ultimately on the fixed effects. In any case, one should compare the fixed-effects estimates and the hypothesis tests regarding the fixed effects across all the estimated models. It is often the case that the testing of fixed effects in highly overparameterized models with degenerate estimates diverge from analogous tests applied to reduced models. In those cases, one should

report the results for the reduced model.

One further alternative for addressing convergence problems is to switch to Bayesian estimation (Gelman et al., 2013), for example as implemented in packages `rstanarm` (function `stan_lmer()`; Gabry & Goodrich, 2016), `blme` (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013), or `MCMCglmm` (Hadfield, 2010). In line with the literature (Bates, Kliegl, et al., 2015; Kimball, Shantz, Eager, & Roy, 2016) we believe that the regularization provided by the priors in a Bayesian framework (as long as the priors are not completely non-informative) is often enough to avoid the problems associated with degenerate or singular fits (e.g., the posteriors of correlation parameters which cannot be identified given a specific data set will simply be extremely wide and include 0, see Bates, Kliegl, et al., 2015, Figure 3). Additionally, the identification of actual convergence problems is comparatively simple via visual inspection of the chains. However, Bayesian approaches require even more care when choosing a contrast scheme, as the prior distribution should ideally be equal for all factor levels, which is not the case for the simple sum-to-zero contrast discussed below (see Rouder, Morey, Speckman, & Province, 2012, p. 363). Furthermore, there is currently no consensus on how to perform hypothesis testing for hierarchical models in a Bayesian framework, especially when a factor contains more than two levels. Consequently, we currently cannot wholeheartedly recommend Bayesian approaches as the default or off-the-shelf procedure for estimating mixed models (but see Singmann, Klauer, & Kellen, 2014). We are hopeful this will change in the next years.

Random Effects Structures for Traditional ANOVA Designs. The estimation of the maximal model is not possible when there is only one observation per participant and cell of a repeated-measures design (i.e., designs typically analyzed using a repeated-measures ANOVA). In this kind of design, the random slopes for the highest-order interaction are perfectly confounded with the residual error term (in more technical language, the model is only identifiable up to the sum of these two variance components). To nevertheless analyze such designs with mixed models the most reasonable decision is to remove the highest-order random slope (e.g., the random slope

for the highest order-interaction). Even though the maximal-random effects structure is not identified in this case, the variability of the random variability of the non-identified effect is added to the residual error term and also incorporated into the standard errors in the appropriate way. We have shown this by simulation elsewhere.[10]. In any case, we strongly recommend researchers to consider this issue before collecting any data. More data and replicates at the level of the cell are always a good idea when estimating mixed models.[11]

Parameterization of Categorical Covariates

All regression-type models, including mixed models, can only be estimated with numerical independent variables. Consequently, factors with m levels need to be transformed into $m - 1$ numerical contrast variables according to a contrast scheme (for an extensive treatment see Cohen, Cohen, West, & Aiken, 2002). Often the choice of contrast scheme does not matter, but there are two situations when it does: For the interpretation of parameters estimates (especially if models include interactions) and for so-called *Type III sums of squares* tests.[12]

The definitions of the different types of sums of squares revolve around whether or not to test lower order effects in the presence (= Type III) or absence (= Type II) of higher order effects. In the statistical literature there is fierce discussion on the "correct" type of sums of squares that we do not want to reiterate here (see Hector, Von Felten, & Schmid, 2010, for an overview). In most cases this only matters for

[10] http://wp.me/p4Y5u1-83

[11] It should be noted again that the discussion in the current paragraph is specifically about a situation in which the *number of replicates* per cells of the design and units of observation (i.e., levels of the random effects grouping factor) is as low as it can be (i.e., 1). This is different from a situation in which the *number of levels* of one random effects grouping factor is low (e.g., 6 or lower). However, in such a situation we recommend treating that effect as fixed (other effects that can be specified as random should remain so).

[12] Please do not confuse the type of sums of squares (here we discuss only Type II and Type III) with the nomenclature used for distinguishing the different types of inferential errors (e.g., Type I and Type II).

unbalanced designs (i.e., for balanced data the different types of sums of squares produce identical results). In the psychological literature on experimental designs (e.g., Howell, 2013; Maxwell & Delaney, 2004) Type III are usually recommended because they are more reasonable under two conditions that are commonly met in psychological experiments: (1) Type III tests assume the imbalance in the data occurs randomly and is not a result of differences in the population (i.e., Type III tests "correct" the imbalance by assuming all cells have the same size). (2) Lower order effects need to be possible in light of interactions (i.e., the pattern of the higher order effect is such that it cancels itself out completely, such as for a perfect cross-over interaction). Consequently, we also recommend to use Type III tests as a default.

A common contrast scheme, which is the default in `R`, is called *treatment contrasts* (i.e., `contr.treatment`; also called dummy coding). With treatment contrasts the first factor level serves as the baseline whereas all other levels are mapped onto exactly one of the contrast variables with a value of 1. As a consequence, the intercept corresponds to the mean of the baseline group and not the grand mean. When fitting models without interactions, this type of contrast has the advantage that the estimates (i.e., the parameters corresponding to the contrast variables) indicate whether there is a difference between the corresponding factor level and the baseline. However, when including interactions, treatment contrasts lead to results that are often difficult to interpret. Whereas the highest-order interaction is unaffected, the lower-order effects (such as main effects) are estimated at the level of the baseline, ultimately yielding what are known as *simple effects* rather than the usually expected lower-order effects. Importantly, this applies to both the resulting parameter estimates of the lower order effects as well as their Type III tests. In other words, a mixed model (or any other regression type model) that includes interactions with factors using treatment contrasts produces parameter estimates as well as Type III tests that often *do not correspond* to what one wants (e.g., main effects are not what is commonly understood as a main effect). Therefore we generally recommend to avoid treatment contrasts for models that include interactions. Note that this issue is independent of whether or not the design is

balanced.

Contrasts schemes that enable an interpretation of both higher- and lower-order effects are *orthogonal* in balanced designs (i.e., the sum of each variable across observations is zero and the sum of the product of all variable pairs is also zero). In such schemes, the intercept corresponds to the grand mean (or the unweighted grand mean in case of unbalanced data) and lower-level effects are estimated at the level of the grand mean. In what follows, we will use one such contrasts known as *effects coding* (i.e., `contr.sum`). In effects coding the last factor level receives a value of -1 on all contrast variables whereas all other factor levels are mapped onto exactly one contrast variable with a value of 1. In the case of a factor with only two levels, the effect-coded parameter value is equal to half of the difference between the two conditions.

One additional complication arises when a regression model includes continuous covariates that interact with other variables included (Cohen et al., 2002). For type III tests with appropriate orthogonal contrasts, the lower-order effects of variables that interact with the continuous covariates are performed at the origin of this covariate (i.e., where it is zero). A common way to deal with this is to center the continuous covariate such that the test of the lower order effect is performed at its mean (e.g., Dalal & Zickar, 2012). However, this might not make sense in all situations. If the zero-point is already meaningful on its own and present in the data, centering is usually not necessary. Another alternative consists of scaling the continuous covariate such that the zero-point becomes meaningful (e.g., to the midpoint of a scale). Note that some authors recommend specific centering schemes for mixed models, mainly for models with nested random effects (e.g., Wang & Maxwell, 2015).

Effect Sizes For Mixed Models

One standardized effect size for categorical fixed effects in mixed models has been developed by Westfall et al. (2014). More specifically, they present formulas for calculating d for mixed models with a single fixed effect with two levels and various random effects structures (see also Judd et al., 2017). In principle, their approach can

also be extended to account for designs with factors with more than two levels and/or interactions (Westfall, 2015, pp. 19-21). However, we are currently not aware of any implementation of this approach.

Fitting Mixed Models in R

The gold standard for fitting mixed models in `R` is function `lmer()` in the `lme4` package (Bates, Mächler, et al., 2015). As for most modeling functions in `R`, the data need to be in a `data.frame` in the *long format* (also known as *tidy data*; Wickham & Grolemund, 2017): Each observation corresponds to one row with one column containing the dependent variable and the other columns containing information pertaining to the identity of the observation (e.g., participant id, item number, conditions). In order to specify the mixed model, `lmer()` needs to be called with a formula and the data as arguments. Table 1 provides an overview of different ways to specify terms in formulas. In the mixed effect formula the dependent variable is on the left side of `~` and the random effects are enclosed in parentheses wherein the pipe symbol "`|`" separates the random effects parameters (on the left) from the random effects grouping factor (on the right). Table 2 provides an overview of different ways for specifying random effects structures. The complete formula for the maximal model described above (i.e., Equation 4) could be:

```
y ~ difficulty + (difficulty|participant) + (difficulty|item)
```

Note that formulas automatically include an intercept, unless it is specifically suppressed via `0` or `-1` (which rarely makes sense).

Using `lmer()` to directly fit a mixed model is associated with the problems already noted above. First, `lmer()` does not provide p-values so that one needs to perform an additional inferential step. Second, the default contrast codes in `R` are such that model with categorical covariates (i.e., `factors`) produce parameter estimates that do not accurately represent lower-order effects (e.g., main effects) if higher-order effects (i.e., interactions) are present. This latter fact is the reason that some people recommend to transform factors into numerical covariates by hand. However, this is not necessary; `R`

Table 1

Model formulas in R

formula	meaning
`a + b`	main effects of `a` and `b` (and no interaction)
`a:b`	only interaction of `a` and `b` (and no main effects)
`a * b`	main effects and interaction of `a` and `b` (expands to: `a + b + a:b`)
`(a+b+c)^2`	main effects and two-way interactions, but no three-way interaction (expands to: `a + b + c + a:b + b:c + a:c`)
`(a+b)*c`	expands to: `a + b + c + a:c + b:c`
`0 + a`	0 suppresses the intercept resulting in a model that has one parameter per level of `a` (identical to: `a - 1`)

Note. Whitespace is ignored in formulas (i.e., `a + b` is identical to `a+b`). The default behavior of `R` is to prevent models with higher order effects in the absence of lower order effects. Thus, `a+a:b` has the same number of parameters as `a*b`, albeit with different parametrization (i.e., `R` will add an additional parameter to `a+a:b` that is not part of the formula).

contains coding schemes that are orthogonal and do not have this problem. The easiest way to change the coding globally is via the `afex` function `set_sum_contrasts()`:

```
require(afex) # this loads afex, equivalent to library(afex)
set_sum_contrasts()
```

Note that the global coding scheme affects *all* `R` modeling functions (e.g., `lm`, `glm`) and not only `lmer()`. But as already mentioned above, for experimental designs orthogonal sum-to-zero contrasts are often a more reasonable default than treatment contrasts.

The `afex` package (Singmann et al., 2017) contains function `mixed()` that is built on top of `lmer()` and avoids both of the problems mentioned in the previous paragraph. It uses orthogonal sum-to-zero contrasts in the default settings and calculates p-values for the terms in the mixed model using the methods discussed above: Kenward-Roger (`method="KR"`, the default), Satterthwaite (`method="S"`), LRTs (`method="LRT"`), and

Table 2

Random effects in `lmer`

random term	meaning
`(1|s)`	random intercepts for `s` (i.e., by-`s` random intercepts), Equation 2
`(1|s) + (1|i)`	by-`s` and by-`i` random intercepts
`(a|s)`	by-`s` random intercepts and by-`s` random slopes for `a` plus their correlation (identical to `(1+a|s)`), Equation 3
`(a*b|s)`	by-`s` random intercepts and by-`s` random slopes for `a`, `b`, and the `a:b` interaction plus correlations among the by-`s` random effects parameters
`(0+a|s)`	by-`s` random slopes for `a` and no random intercept
`(a||s)`	by-`s` random intercepts and by-`s` random slopes for `a`, but no correlation (expands to: `(0+a|s) + (1|s)`)

Note. Suppressing the correlation parameters via `||` works only for numerical covariates in `lmer` and not for factors. `afex` provides the functionality to suppress the correlation also among factors if argument `expand_re = TRUE` in the call to `mixed()` (see also function `lmer_alt()`).

parametric bootstrap (`method="PB"`).[13] An example analysis could be (assuming the data is in a `data.frame` called `d`):

```
require(afex)
```

[13]Contrary to the description given above, *p*-values for the first two methods are calculated directly from the full model using Wald tests in which the parameters corresponding to the specific effects are set to 0 (Fox, 2015). This is faster than explicitly comparing nested models, but equivalent for those two methods. For the latter two methods, the *p*-values are calculated via the comparison of nested models; `mixed()` fits an encompassing model with all parameters and one reduced model corresponding to each of the model terms in which the parameters corresponding to the term are withhold from the full model (all fits are performed with `lmer()`). Estimating the different models can be distributed across different CPU cores to speed up the overall fitting process. After estimating all necessary models the *p*-values are calculated with the corresponding method.

```
set_sum_contrasts() # not strictly necessary, but always a good idea
m1 <- mixed(y ~ difficulty + (difficulty|participant) + (difficulty|item), d)
m1
```

Note that loading `afex` also loads `lme4`, but at the same time replaces the original `lme4` `lmer()` function with a function of the same name from package `lmerTest` (Kuznetsova et al., 2016). The only difference between the `lmerTest` versions of `lmer()` compared to the original `lme4` version is that the output of the former includes p-values based on the Satterthwaite approximation in the standard output. Furthermore, the result can be passed to the `anova` function which then reports p-values for the effects. Note again, that one needs to be careful and use appropriate (sum-to-zero) contrasts whenever `lmer()` is invoked directly and parameters estimated interpreted or the type III tests calculated.

When fitting mixed models with complicated random effects structures, convergence warnings appear frequently. It is important to distinguish between the warnings that indicate that the results are not to be trusted and those that suggest there is a problem, but the results may still be interpretable. All warnings from the optimization algorithm (the default one is `bobyqa`) belong into the former category. One such message notes that the `convergence code` is `1` and/or specifies that the "`maximum number of function evaluations` [was] `exceeded`". One way to address this warning is to increase the maximum number of function evaluations and rerun the model. The following code increases the maximum number of iterations to 1 million:

```
mixed(..., control = lmerControl(optCtrl = list(maxfun = 1e6)))
```

Note that the same argument could also be used in the call to `lmer()`. A warning unique to `mixed()` indicating that the results cannot be trusted is reported if a nested model provides a better fit than a superordinate (i.e., the full) model. As such a pattern is logically impossible, such a warning indicates the presence of a severe problem caused by the present combination of data, model, and optimization algorithm. One way to address this problem is to change the optimization algorithm or trying a variety of

optimization algorithms (by setting `all_fit=TRUE`). However, it is our experience that changing the optimizer sometimes does not always solve this kind of problem, as the warning may indicate that there is too few data for estimating the desired random-effects structure (Bates, Kliegl, et al., 2015).

`lmer()` also performs a variety of convergence checks on the obtained results that may result in a warning (indicated by '`In checkConv(...)`'). These warnings do not necessarily indicate that the results cannot be trusted. It is possible that these warnings are false positives and the model converged successfully. In this case, the warnings can be safely ignored. Another possible reason for these warnings is that some of the (usually random effects) parameters have identifiability or other numerical problems and the fit is singular or degenerate. However, this does not necessarily affect the tests of fixed effects (but see Matuschek et al., 2017). As mentioned above, variance estimates of 0 and correlations at the boundary are clear indications of degenerate parameter estimates. In this case one could try to refit the model without the problematic random effects parameters. A good strategy is often to start by removing the correlation among the random effects parameters. `mixed` allows the user to suppress the correlation for random effects (even for factors) if `expand_re=TRUE`. When setting `expand_re=TRUE`, the random effects factors are transformed into numerical covariates for which the correlation parameters are then suppressed. For example, the following code suppresses the correlations among the by-`id` random effects parameters, but not among the by-`item` random effects parameters:

```
mixed(y~x1*x2+(x1*x2||id)+(x1|item), expand_re=TRUE)
```

Some more advice on how to address and diagnose convergence warnings are provided by `lme4` author Ben Bolker[14] and in Bates, Kliegl, et al. (2015). Again, it is important to keep in mind that there is no guarantee that a given data set converges for a desired random effects structure as random effects parameters are more difficult to estimate than fixed effects parameters.[15]

[14] see https://rpubs.com/bbolker/lme4_convergence and https://rpubs.com/bbolker/lme4trouble1

[15] "The combination of some data and an aching desire for an answer does not ensure that a reasonable

One final note regarding `lmer` and `mixed` is that models fitted with both functions can be directly used for follow-up tests, planned contrasts, or plotting with the methods implemented in package `lsmeans` (Lenth, 2015) or `effects` (Fox, 2003). Importantly, these methods take the dependencies in the data into account. A comprehensive example analysis for a factorial design with crossed-random effects that also includes follow up-tests is provided at https://cran.r-project.org/package=afex/vignettes/afex_mixed_example.html.

Beyond Linear Mixed Models and the Identity Link Function

All models discussed in this chapter so far share one commonality, namely the assumption that the residuals follow a normal distribution with variance σ_ϵ^2. This is the normality assumption that linear mixed models share with ordinary linear regression and ANOVA models. However, not all data can be reasonably described under such an assumption. For example, a common dependent variable in experimental designs is accuracy (i.e., one binary response per trial) which follows a binomial distribution and should therefore not be analyzed with models assuming a normal distribution (Dixon, 2008). One adequate procedure to analyze categorical data like the one following a binomial distribution is *logistic regression*, which is a member of the class of *generalized linear models* (GLMs; McCullagh & Nelder, 1994). GLMs differ from ordinary linear models by allowing the specification of the *residual distribution* and the *link function* that maps the model predictions – which are defined on a real scale – onto the (dependent variable's) outcome space. In the case of logistic regression, the residual distribution is binomial and the link function the logistic function. Another GLM for binomial data that is common in economics is probit regression. It again assumes a binomial residual distribution but instead uses the probit function (i.e., the cumulative distribution function of the normal distribution) as a the link between the model predictions and the dependent variable's outcome space. Other residual distributions such as the Poisson (suitable for handling count data) are also possible. However, unlike

answer can be extracted from a given body of data." (Tukey, 1986, pp. 74)

ordinary linear regression – which can be seen as a special case of GLMs with normal residual distribution and identity link function – there are no repeated-measures variants of GLMs. Hence, standard GLMs cannot account for non-independence due to repeated measures and cannot be used for within-subject designs.

Fortunately, linear mixed models can be extended to *generalized linear mixed models* (GLMMs) that also allow to specify the residual distribution and link function, but also allow for the inclusion of random effects. Several of the issues discussed for LMMs (such as the specification of random effects and factor codings) apply in exactly the same way to GLMMs. Furthermore, GLMMs can also be estimated with function `mixed` by passing a family argument and an appropriate method for testing such as LRT (e.g., `mixed(..., family=binomial(link="logit"), method = "LRT")`). However, due to the nonlinear nature of most link functions, the interpretations of most model predictions, specifically of lower-order effects in factorial designs, can be quite challenging. Additionally, specifically binomial GLMMs can be quite prone to producing singular fits or other convergence problems due to the limited amount of information provided by each data point (i.e., 0 or 1; see Eager & Roy, 2017). A comprehensive introduction to GLMMs is beyond the scope of the current chapter so we refer the interested reader to further literature on this matter (e.g., Bolker, 2015; Jaeger, 2008).

Summary

Mixed models are a modern class of statistical models that extend regular regression models by including random effects parameters to account for dependencies among related data points. More specifically, these random effects parameters control for the stochastic variability associated with the levels of a random effects grouping factor (e.g., participant, item, group) by adjusting the fixed effects parameters with idiosyncratic displacements or offsets. This essentially gives each level of the random grouping factor its own set of regression parameters under the restriction of hierarchical shrinkage implementing an efficient data analysis strategy also known as *partial pooling*. Importantly, modern mixed model implementations allow to simultaneously control for

multiple independent (i.e., crossed) sources of stochastic variability. The goal of this chapter was to provide a general introduction to the concepts underlying mixed models and to walk through the steps necessary to set up a mixed model in `R` with a special focus on common hurdles researchers may encounter. To reiterate, the most important step for researchers is to identify the maximal random effects structure which is given by random intercepts for each source of random variation (i.e., random effects grouping factor), random slopes for all fixed effects that vary within the levels of a random effects parameter, and correlations among all random effects parameters for a given grouping factor. Once the appropriate random effects structure for a design has been identified, tests of fixed effect (assuming appropriate contrasts) take the stochastic variability into account, but can be interpreted as tests of effects in a regular ANOVA. For more information on specifying the random effects structure we recommend Barr et al. (2013). Aspects of mixed models that go beyond the issues discussed here are given in Bolker (2015) and Snijders and Bosker (2012). An even more powerful class of models than discussed here, *generalized additive mixed models* (GAMMs), is described in Baayen, Vasishth, Kliegl, and Bates (2017).

References

Baayen, H. (2008). *Analyzing linguistic data : a practical introduction to statistics using r*. Cambridge, UK; New York: Cambridge University Press.

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi:10.1016/j.jml.2007.12.005

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. doi:10.1016/j.jml.2016.11.006

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D. (2006, May). [r] lmer, p-values and all that [electronic mailing list message]. Retrieved from https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html.

Bates, D. & DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*. Special Issue on Semiparametric and Nonparametric Mixed Models, *91*(1), 1–17. doi:10.1016/j.jmva.2004.04.013

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv:1506.04967 [stat]*. arXiv: 1506.04967.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). doi:10.18637/jss.v067.i01

Bolker, B. M. (2015). Linear and generalized linear mixed models. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), *Ecological statistics: contemporary theory and application* (pp. 309–333).

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*(4), 685–709. doi:10.1007/s11336-013-9328-2

Clark, H. H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, *12*(4), 335–359. doi:10.1016/S0022-5371(73)80014-3

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied multiple Regression/Correlation analysis for the behavioral sciences*. New York: Routledge Academic.

Dalal, D. K. & Zickar, M. J. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods*, *15*(3), 339–362. doi:10.1177/1094428111430540

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456. doi:10.1016/j.jml.2007.11.004

Eager, C. & Roy, J. (2017). Mixed effects models are sometimes terrible. *arXiv:1701.04858 [stat]*. arXiv: 1701.04858.

Fox, J. (2003). Effect displays in r for generalised linear models. *Journal of Statistical Software*, *8*(15), 1–27.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Los Angeles: Sage.

Gabry, J. & Goodrich, B. (2016). *Rstanarm: bayesian applied regression modeling via stan*. R package version 2.13.1.

Gelman, A. (2005). Analysis of variance: why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–31. doi:10.2307/3448650

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis, third edition*. Hoboken: CRC Press.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the **MCMCglmm** *R* package. *Journal of Statistical Software*, *33*(2). doi:10.18637/jss.v033.i02

Halekoh, U. & Højsgaard, S. (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models – the r package pbkrtest. *Journal of Statistical Software*, *59*(9).

Hector, A., Von Felten, S., & Schmid, B. (2010). Analysis of variance with unbalanced data: an update for ecology & evolution. *Journal of Animal Ecology*, *79*(2), 308–316. doi:10.1111/j.1365-2656.2009.01634.x

Howell, D. C. (2013). *Statistical methods for psychology*. Belmont, CA: Wadsworth Cengage Learning.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. doi:10.1016/j.jml.2007.11.007

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. doi:10.1037/a0028347

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601–625. doi:10.1146/annurev-psych-122414-033702

Kenny, D. A. & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*(3), 422–431. doi:10.1037/0033-2909.99.3.422

Kenward, M. G. & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997. doi:10.2307/2533558

Kimball, A., Shantz, K., Eager, C., & Roy, J. (2016). Beyond maximal random effects for logistic regression: moving past convergence problems. *arXiv:1611.00083 [stat]*. arXiv: 1611.00083.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: tests in linear mixed effects models*. R package version 2.0-33.

Lenth, R. V. (2015). *Lsmeans: least-squares means*. R package version 2.20-1.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type i error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. doi:10.1016/j.jml.2017.01.001

Maxwell, S. E. & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model-comparisons perspective*. Mahwah, N.J.: Lawrence Erlbaum Associates.

McCullagh, P. & Nelder, J. A. (1994). *Generalized linear models*. London: Chapman & Hall.

Pinheiro, J. C. & Bates, D. (2000). *Mixed-effects models in s and s-PLUS*. New York: Springer.

R Core Team. (2016). *R: a language and environment for statistical computing*. http://www.R-project.org/. Vienna, Austria: R Foundation for Statistical Computing.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. doi:10.1016/j.jmp.2012.08.001

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*(5), 309–316. doi:10.1007/BF02288586

Scheipl, F., Greven, S., & Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational statistics & data analysis*, *52*(7), 3283–3299.

Schielzeth, H. & Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420. doi:10.1093/beheco/arn145

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2017). *Afex: analysis of factorial experiments*. R package version 0.17-8. http://cran.r-project.org/package=afex.

Singmann, H., Klauer, K. C., & Kellen, D. (2014). Intuitive logic revisited: new data and a bayesian mixed model meta-analysis. *PLoS ONE*, *9*(4), e94223. doi:10.1371/journal.pone.0094223

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

Tukey, J. W. (1986). Sunset salvo. *The American Statistician*, *40*(1), 72–76. doi:10.1080/00031305.1986.10475361

Wang, L. & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods*, *20*(1), 63–83. doi:10.1037/met0000030

Westfall, J. (2015). PANGEA: power ANalysis for GEneral anova designs. *Unpublished manuscript. Available at http://jakewestfall. org/publications/pangea. pdf.*

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. doi:10.1037/xge0000014

Wickham, H. & Grolemund, G. (2017). *R for data science: import, tidy, transform, visualize, and model data*. OCLC: 927922708. Sebastopol CA: O'Reilly.