An approachable introduction to linear mixed effects modeling with implementation in R

Violet A. Brown

Washington University in St. Louis

Contact: violet.brown@wustl.edu

Abstract

This tutorial serves as both an approachable theoretical introduction to mixed effects modeling and a practical introduction to how to implement these models in R. The intended audience is any researcher who has some basic statistical knowledge, but little or no experience implementing mixed effects models in R using their own data. In an attempt to increase the accessibility of this paper, I deliberately avoid using mathematical terminology beyond what a student would learn in a standard graduate-level statistics course, but I reference articles and textbooks that provide more detail for interested readers. This tutorial includes snippets of R code throughout, as well as the data and R script used to build the models described in the text so readers can follow along if they wish. The goal of this practical introduction is to provide researchers with the tools they need to begin implementing mixed models in their own research.

An approachable introduction to linear mixed effects modeling with implementation in R

Background

In many areas of experimental psychology, researchers collect data from participants responding to multiple trials. This type of data has traditionally been analyzed using repeated measures analyses of variance (ANOVAs)—statistical analyses that assess whether conditions differ significantly in their means, accounting for the fact that observations within individuals are correlated. Repeated measures ANOVAs have been favored for analyzing this type of data because using other statistical techniques, like ordinary least squares regression, would violate a crucial assumption of many statistical tests: the *independence assumption*. This assumption states that the observations in a dataset must be independent; that is, they cannot be correlated with one another. But take, for example, a reaction time study in which participants respond to the same 100 trials, each of which corresponds to a different item (e.g., a particular word in a psycholinguistics study). Here, reaction times within a given participant and within an item will certainly be correlated—some participants are faster than others, and some items are responded to more quickly than others. Given that observations are not independent, data in which participants respond to multiple trials must be analyzed with a statistical test that takes the dependencies in the data into account.

For this reason, repeated measures ANOVAs are preferable to standard ANOVAs and ordinary regression for analyzing data in which observations are nested within participants; both alternatives ignore the hierarchical structure of the data. However, repeated measures ANOVAs are far from perfect. Although they can model either participant- or item-level variability (often referred to as F1 and F2 analyses in the ANOVA literature), they cannot simultaneously take both sources of variability into account, so observations within a condition must be collapsed across either items or participants. When we aggregate in this way, however, important information about variability within participants or items is lost, which reduces statistical power—the likelihood of detecting an effect if one exists (see Barr, 2008). Further, repeated measures ANOVAs assume that trial spacing is consistent across participants, so this technique is inappropriate in situations in which trial spacing varies from one participant to another, as is the case in many longitudinal designs (e.g., Schober & Vetter, 2018).

Another issue with ANOVAs is that they deal with missing cases via *listwise deletion*—this means that if a single observation within an individual is missing, the entire case is deleted and none of the observations from that individual will be used in the analysis. Depending on the number of complete cases in the dataset, this can substantially reduce sample size, leading to inflated standard error estimates and reduced statistical power (though the estimates will be unbiased if the data are missing completely at random, see Enders, 2010). ANOVAs also assume that the dependent variable is continuous and the independent variables are categorical; continuous predictors (e.g., time in a longitudinal study) must therefore be treated categorically (i.e., binned), which reduces statistical power and makes it difficult to model non-linear relationships between predictors and outcomes (e.g., Liben-Nowell et al., 2019; Royston et al., 2005). A final drawback of ANOVAs is that although they indicate whether an effect is significant, they do not provide information about the magnitude or direction of the effect; that is, they do not provide individual coefficient estimates for each predictor that indicate growth or trajectory.

Mixed Effects Models Take the Stage

The shortcomings of ANOVAs and ordinary least squares regression described above can be avoided by using linear mixed effects modeling (also referred to as hierarchical linear regression, multilevel modeling, or mixed modeling). Mixed effects modeling allows a researcher to examine the condition of interest while also taking into account variability within and across participants and items simultaneously. It does not assume equal trial spacing across participants, and it handles missing data quite well—put simply, missing cases are dropped, but participants or items with more missing cases have weaker influences on parameter estimates. Further, given that mixed effects modeling is a type of linear regression, continuous predictors do not pose a problem to the analysis, and the fitted model provides coefficient estimates that indicate the average relationship between each predictor and the outcome, controlling for all other predictors in the model. Mixed effects modeling is therefore appropriate in many cases in which standard ANOVAs, repeated measures ANOVAs, and ordinary least squares regression are not, making it a more flexible analytical tool.

Introducing the Data

This paper uses examples from my own research area on human speech perception, but the concepts apply to a wide variety of areas within and beyond psychology. For example, participants might be presented with multiple items to be recalled in a memory task, view videos and be asked to evaluate affect associated with each of them in a social psychology experiment, or read a series of narratives and be asked to describe the extent to which each of them generates anxiety in a clinical experiment.[1] The goal of this paper is to provide a practical introduction to linear mixed effects modeling and introduce the tools that will enable you to conduct these analyses on your own. This overview is not intended to address every issue you may encounter in your own analyses, but is meant to provide enough information that you have a sense of what to ask if you get stuck. To help you along the way, I will provide snippets of R code using dummy data that will serve as a running example throughout the paper.

The example data I provide (see https://osf.io/v6qag/), which we will work with in the "Examples and Implementation in R" section, comes from a hypothetical within-subjects response time study in which each of 65 participants in a speech perception study were presented 398 words, some in the auditory modality alone (audio-only; coded 0) and some with an accompanying visual stimulus (audiovisual; coded 1). Though these data are not real, you can think of the visual stimulus as the face of the talker that produced the speech. I will use mixed effects modeling to assess the effect of modality

[1] It is important to note that the examples in this paper concern *crossed* rather than *nested* random effects. Random effects (defined below) such as participants and items are considered crossed when every participant responds to every item, and nested when every participant responds to a different set of items. The classic example of nested random effects comes from education research in which students are nested within classes, which in turn are nested within schools (see Raudenbush, 1988). The motivation for using mixed modeling applies to both design types, but the examples and R code I provide assume a crossed design (see Baayen et al., 2008; Judd et al., 2017; Quené & van den Bergh, 2008; Westfall et al., 2014 for more on the distinction between crossed and nested designs).

(audio-only versus audiovisual) on response times (and word intelligibility later in the paper) while simultaneously modeling variability both within and across participants and items.

The left panel of Table 1 shows the first six lines of the data in the desired format: *unaggregated long format*. If you are following along with your own data, before moving forward you should make sure your data are in long format such that each row represents an individual observation (i.e., do not aggregate across either participants or items). Notice that the first six rows each correspond to a different word ("stim") presented to the same participant ("PID"). In contrast, if we were running an ANOVA the data frame would likely contain two rows per participant, one for each modality, and the value in the response time ("RT") column would reflect the mean response time for all words presented to that individual in each condition (right panel of Table 1).

Table 1. First six rows of a dataset in the desired unaggregated format (left) compared to the first six rows of the same dataset that was aggregated across items (right).

```
PID    modality  stim   RT                PID    modality        RT
201 Audiovisual  unit  792                201  Audio-only  1117.528
201 Audiovisual  week  893                201  Audiovisual 1261.060
201 Audiovisual  dump  890                202  Audio-only  1209.327
201 Audiovisual candy  887                202  Audiovisual 1303.745
201 Audiovisual sweat 1341                203  Audio-only  1899.286
201 Audiovisual  boss 1303                203  Audiovisual 1885.930
```

Fixed and Random Effects

Mixed effects models are called "mixed" because they simultaneously model *fixed* and *random* effects. Though consistent definitions are difficult to come by, typically an effect is considered *fixed* if all levels of interest are included in the study and the influence of the effect is common across participants and items. An effect is considered *random* if the levels represent a random sample from the population of interest and the way the effect operates varies across participants (e.g., Snijders & Bosker, 2012). In our hypothetical experiment, words and participants are modeled as random effects because there are more than 398 words and 65 people in the respective populations of interest, and our words and participants simply represent random samples from those populations. Modality is included as a fixed effect because we expect that there is a common relationship between modality and response times that applies to all participants and items.

Including random effects for participants and items resolves the non-independence problem that often plagues ordinary regression by accounting for the fact that some participants respond more quickly than others, and some items are responded to more quickly than others. These random deviations from the mean response time are called *random intercepts*. For example, the model may estimate that the mean response time for some condition is 850 milliseconds (ms), but specifying by-participant random intercepts allows the model to estimate each participant's deviation from this fixed estimate of the mean response time. So if one participant tends to respond particularly quickly, their individual intercept might be shifted down 150 ms. Similarly, including by-item random intercepts enables the model to estimate each item's deviation from the fixed intercept, reflecting the fact that some words tend to be responded to more quickly than others. In ordinary regression, however, the same regression line is applied to all participants and items, so predictions tend to be less accurate and residual error tends to be larger in standard regression than in mixed effects regression. Thus, in mixed modeling the fixed intercept estimate essentially represents the average intercept, and random intercepts allow each participant and item to deviate from this average.[2] These deviations (and all random effects) are assumed to follow a normal distribution with a mean of zero and a variance that is estimated by the model.

In addition to random intercepts, another source of variability that mixed effects models can account for comes from the fact that a variable that is modeled as a fixed effect may actually have different influences on different participants (or items). In our example, some participants may show very

[2] Note that this is not literally how parameters in mixed effects models are estimated. Those details are beyond the scope of this paper, and this simplified description is provided to help you conceptualize what mixed models are doing behind the scenes. See Snijders and Bosker (2012) for more detail.

small differences in response times between the audio-only and audiovisual conditions, and others may show large differences. Similarly, some words may be more affected by modality than others. To model this type of variability, we would include *random slopes* in the model specification. In our hypothetical study, the model may estimate that the main effect of modality is 96 ms—meaning that participants are, on average, 96 ms slower in the audiovisual condition than the audio-only condition—but one participant may be very strongly affected by modality (e.g., a response time difference between modalities of 200 ms) and another may be only weakly affected by modality (e.g., a response time difference between modalities of 10 ms). These individual deviations from the average modality effect are modeled via random slopes (note that a simple mean difference like the one in this hypothetical experiment is represented in a regression equation as a slope).

It may be confusing that modality is modeled as both a fixed and a random effect, but recall that an effect is considered fixed if all levels of interest are included in the study and the effect is assumed to influence participants in a common way. In our case, modality is fixed because the two levels of the factor (audio-only and audiovisual) are the only levels we are interested in—that is, we have exhausted all levels of this particular population—and we are modeling the common influence of modality on response times across participants and items. However, given that participants represent a random sample from the population of interest and we have *not* exhausted all levels of that population, the effect of modality within participants represents a subset of possible ways modality and participants can interact. In other words, modality itself is not the random effect, but rather the way it interacts with participants is random—including the random slope for modality allows the model to estimate each participant's deviation from the overall (fixed) trend. For more on the distinction between fixed and random effects and a description of when a researcher may actually want to model participants as a fixed effect, see Mirman (2014).

Visualizing the Influence of Random Effects

In this section, I provide plots that will help you visualize what happens when you build on ordinary regression by introducing random intercepts and random slopes. These plots are derived from fake data from four hypothetical participants each responding to four items, and we are interested in the influence of word difficulty (where 0 represents “very easy” by some collection of criteria, like the frequency with which the word occurs in the language and the number of similar sounding words, and 10 represents “very difficult”) on response times. First consider a model with no random effects (i.e., ordinary least squares regression; Figure 1). More difficult words tend to elicit slower response times, but because there are no random effects, the model estimates are the same for every participant and item; that is, you (and indeed the model) cannot tell which points in the plot correspond to which participants or items. Further, given that this model predicts just one regression line that applies to all observations, the residual error (represented by vertical lines connecting each point to the regression line) is relatively large.

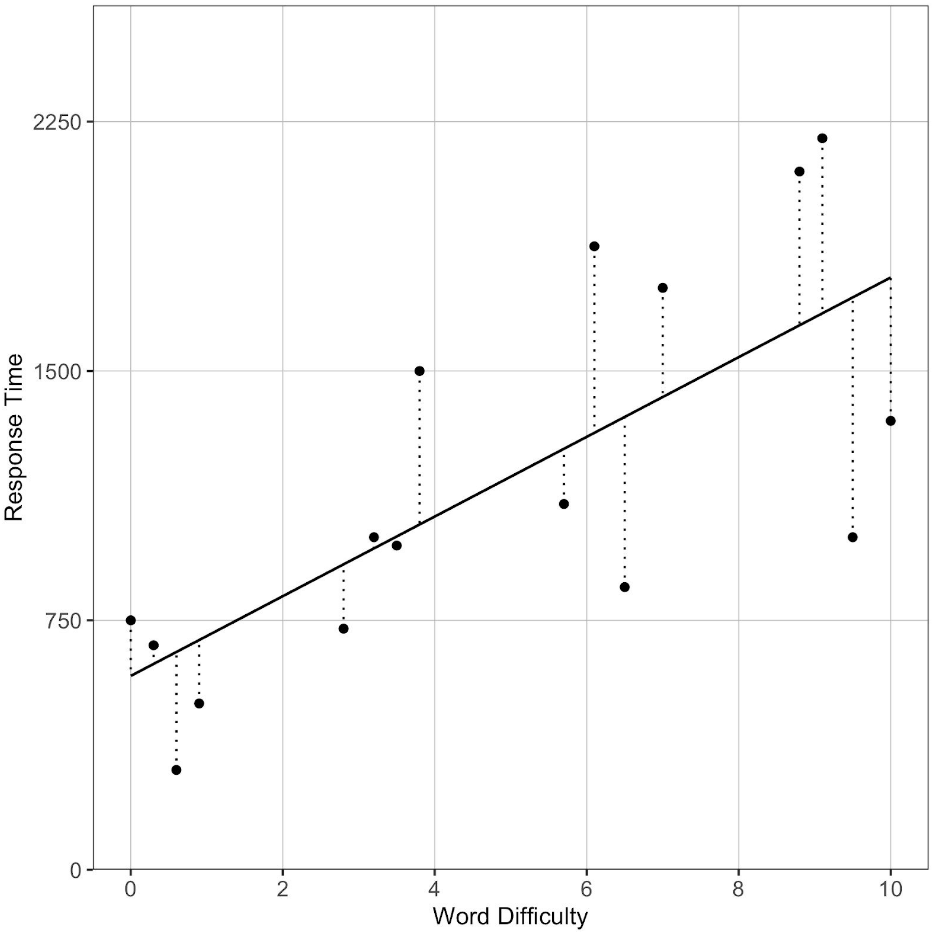

Figure 1. Ordinary least squares regression line depicting the relationship between word difficulty and response time for fake data. This model includes only fixed effects. Each dot represents a single observation of one participant responding to one word, and the vertical lines represent deviation of each point from the line of best fit (i.e., error). Note that the nested nature of the data cannot be discerned from this plot; indeed, ordinary least squares regression does not take dependencies in the data into account.

Next consider a model that includes a random intercept for participants. In Figure 2, each dashed line depicts model predictions for a single participant, and the solid line depicts the estimates for the fixed effects. This model takes into account the fact that some participants tend to have slower response times than others. Here, the overall effect of word difficulty on response times is still apparent, but this model does a better job predicting response times for a given participant because it allows for each participant to have a different intercept (representing the predicted response time for a word with a 0 on the difficulty scale). In this example, the relationship between word difficulty and response time is equally strong for all participants (i.e., the slope is fixed); random intercepts simply shift each participant's regression line up or down depending on that individual's deviation from the mean. Notice that the residual error is substantially smaller in the random intercepts model relative to the ordinary least squares model. This is because we have taken into account the fact that some participants have faster response times than others (i.e., each participant's intercept can vary from the mean intercept), so residual error represents deviation from a specific participant's regression line rather than the overall regression line.

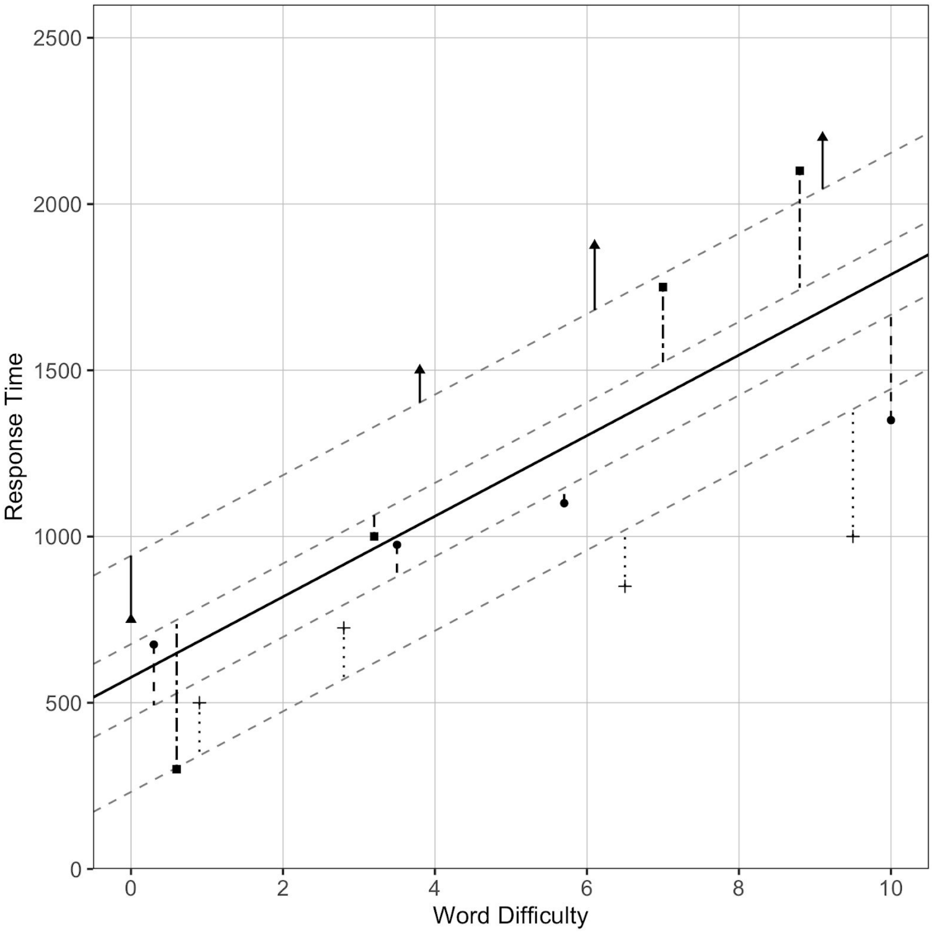

Figure 2. Mixed effects regression lines depicting the relationship between word difficulty and response time for fake data, generated from a model including by-participant random intercepts but no random slopes. Each dashed line represents model predictions for a single participant, and the solid line represents the fixed effects estimates for the intercept and slope. The shapes represent individual observations of each word for a given participant, and the vertical lines represent the deviation of each data point from the participant's individual regression line. Notice that including random intercepts reduces residual error relative to error in the ordinary least squares model.

Figure 3 shows how the model changes when we include by-participant random slopes; this model allows for the relationship between word difficulty and response time to vary across participants. Here, not only do participants differ in how fast they tend to respond (reflected in random intercepts), but they also differ in the extent to which they are affected by changes in word difficulty (reflected in random slopes). Although the general trend that difficult words are responded to more slowly is still apparent, the strength of this relationship varies across participants, and participants have different estimated intercepts. The result is that the residual error is much smaller because each regression line is tailored to the individual.

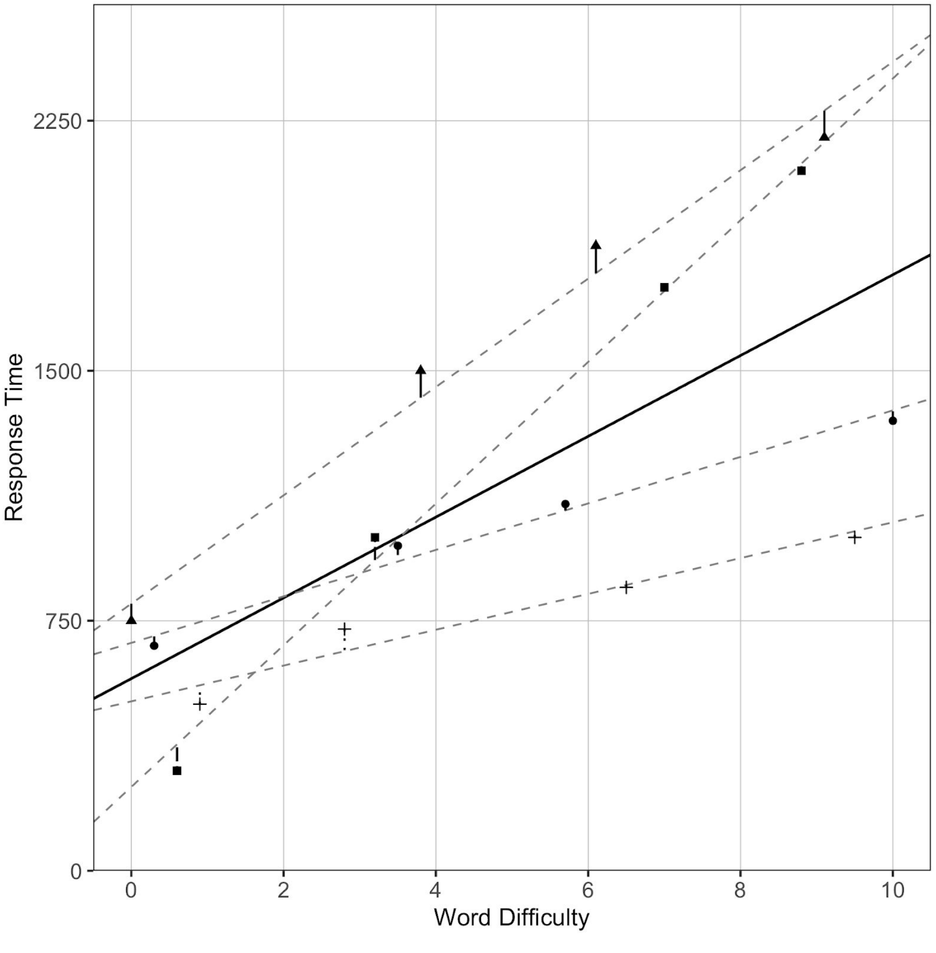

Figure 3. Mixed effects regression lines depicting the relationship between word difficulty and response time for fake data, generated from a model including by-participant random intercepts as well as by-participant random slopes for word difficulty. Each dashed line represents model predictions for a single participant, and the solid line represents the fixed effects estimates for the intercept and slope. The shapes represent individual observations of each word for a given participant, and the vertical lines

represent the deviation of each data point from the participant's individual regression line. Notice that including random slopes reduces residual error relative to error in both the random intercepts model and the ordinary least squares model.

Which Random Effects Should You Include?

Before we move on to implementation in R, it is important to note one other issue regarding random effects structures in mixed effects modeling: deciding which random slopes are justified by your design. In the example above, the question of interest is whether response times to words differ depending on their difficulty. Word difficulty is manipulated within-subjects, but because the words differ on an intrinsic property—namely, their difficulty—word difficulty is a between-items variable. Given that by-item random slopes account for variability across items in the extent to which they are affected by the independent variable of interest, we cannot model the effect of word difficulty on a particular item because each word has only one level of difficulty. By-participant and by-item slopes are only justified for within-subjects and within-items designs, respectively.[3] Thus, our random effects structure can include random intercepts for both participants and items, as well as by-participant random slopes for word difficulty, but *cannot* include by-item random slopes for word difficulty. Similarly, if we suspected that people with better hearing might be less affected by word difficulty, we could include hearing ability as a fixed effect, but we could *not* include by-participant random slopes for hearing ability. This is because each participant would have only one value in the dataset representing their hearing ability, so we cannot model the effect of hearing ability within participants. Because including by-item random slopes for word difficulty and by-participant random slopes for hearing ability would not be justified in this example, the random effects structure described would represent the *maximal random effects structure justified by the design* (see Barr et al., 2013). In cases in which by-participant and by-item random slopes are justified,

[3] Note that it is possible that including by-item random slopes might be unjustified even when the conditions are not defined by stimulus-intrinsic properties. For example, if you are interested in the effect of background noise on response times to words, but different words are assigned to different conditions (each word appears in only one condition), it would not be justified to include by-item random slopes. It is therefore crucial to consider your experimental design before building mixed effects models.

mixed effects models can incorporate the simultaneous influences of both participant and item random slopes (see Barr et al., 2013 for a helpful visualization depicting the simultaneous influences of participant and item random effects).

Examples and Implementation in R

In order to run a mixed effects model in R, install R (R Core Team, 2016) and then RStudio, a programming environment that allows you to write R code, run it, and view graphs and data frames all in one place. I suggest working in RStudio rather than R (this is not a rule—some people code in the R console without RStudio, but those people are wizards). Links to install these programs are available in the "Helpful Links" section at the end of this article. Base R (the set of tools that is built into R) has a host of functions, but to create mixed effects models you will need to install a specific package called *lme4* (Bates et al., 2014). Packages, also referred to as libraries, are sets of functions that work together that are not already built into Base R. To install a package, run the following line of code (you should only run this line of code if you have not already installed the package):

```
> install.packages(“lme4”)
```

Once the package is installed, it is always on your computer and you will not need to run that line of code again. Whenever you want to create mixed effects models you will need to load the installed package, which will give you access to all the functions you need (but you need to re-run this line of code every time you start a new R session):

```
> library(lme4)
```

Analyzing data with a continuous outcome (response time)

Now we can start building some models. For these examples, I will be using the dummy data described above, which comes from a hypothetical study assessing whether a visual stimulus affects response times and word intelligibility. To follow along, go to https://osf.io/v6qag/ and navigate to the R script called "intro_to_lmer.R." In this hypothetical study, assume that modality was manipulated within-subjects and within-items, meaning each word was presented in audio-only and audiovisual conditions, and each participant was presented with both modalities (but each word only occurred in one modality for each participant). Thus, the maximal random effects structure justified by the design includes random intercepts for participants and items, as well as by-participant and by-item random slopes for modality, meaning that both participants and words might differ in the extent to which they are affected by modality.

In this example I used a dummy coding (also referred to as treatment coding) scheme such that the audio-only condition served as the reference level and was therefore coded as 0, and the audiovisual condition was coded as 1. Thus, when we view the output from the mixed effects models, the regression coefficient associated with the intercept represents the estimated mean response time in the audio-only condition (when modality = 0), and the coefficient associated with the effect of modality indicates how the mean response time changes when we move from the audio-only to the audiovisual condition (when modality = 1). We could have used the audiovisual condition as the reference level, in which case the intercept would represent the estimated mean response time in the audiovisual condition (when modality = 0), and the modality effect would indicate how this estimate changes when we move from the audiovisual condition to the audio-only condition (when modality = 1). Changing the coding scheme, either by changing the reference level or changing to a different coding scheme altogether (e.g., sum/deviation coding, which involves coding the groups as -0.5 and 0.5 or -1 and 1 so the intercept corresponds to the grand mean) will not change the fit of the model; it will simply change the interpretation of the regression coefficients. See Wendorf (2004) for a helpful description of various coding schemes.

For reference, the basic syntax for mixed effects modeling for an experiment with one independent variable and random intercepts but no random slopes for (crossed) participants and items is:

```
> lmer(outcome ~ 1 + predictor + (1 | participant) + (1 | item), data = data)
```

However, in our example, we want to include both random intercepts and random slopes for participants and items, which requires the following syntax:

```
> lmer(outcome ~ 1 + predictor + (1 + predictor | participant) + (1 +
predictor | item), data = data)
```

The portions in parentheses are the random effects, and those not in parentheses are the fixed effects. The `lmer` part is the function that builds a mixed effects model (which you can access because you installed the *lme4* package). In plain language, this syntax means "please predict the outcome from the predictor and the random intercepts and slopes for participants and items, using the data I provide."[4]

The 1 outside the parentheses indicates that the model should estimate a fixed intercept, and the 1 inside parentheses indicates that the model should estimate a random intercept for whichever grouping factor appears to the right of the vertical line. Adding "`+ predictor`" to the portion in parentheses indicates that you want to include a random slope in addition to the intercept. Note that including the 1 is optional in both the fixed and random effects portions; I typically include it for clarity,[5] but it is not strictly necessary. In the example we will be working with, the full model (i.e., the model including the fixed effects of interest and all random effects) is specified as follows:

```
> rt_full.mod <- lmer(RT ~ 1 + modality + (1 + modality | PID) + (1 + modality
| stim), data = rt_data)
```

Here, we are predicting response times (`RT`) based on the fixed effects for the intercept and `modality` (audio-only versus audiovisual), we are including random intercepts and slopes for both participants (`PID` = participant identification number) and words (`stim` = stimuli), and we are telling R to

[4] When you are creating mixed effects models like these, what R is doing is computing the values of the parameters that maximize the likelihood of the data given the structure that you specify for the model using *maximum likelihood estimation* (see Etz, 2018 for an approachable introduction to the concept of likelihood).

[5] In addition to providing clarity, including the 1 in the model specification as a placeholder for the intercept also serves as a helpful reminder that the intercept can be explicitly excluded, as in no-intercept models, and as a reminder that replacing it with a 0 in the random effects portion can allow us to model random intercepts and slopes independently. No-intercept models are beyond the scope of this paper, but we will briefly discuss independently modeling random intercepts and slopes below.

use the data frame called `rt_data`.[6] Also note that the above line of code includes a symbol that I have not introduced yet: the `<-` operator. This is used to assign a name to an object (a data structure with specific attributes that is stored in R's memory) and save it for later. Here, we created a model and gave it an intuitive name so that we know what that object represents later on.

If you are running that line of code in the R script, you may notice that you get a warning message saying that the model failed to converge. These models are computationally complex, especially when they have rich random effects structures, and failure to converge basically means that a good fit for the data could not be found within a reasonable number of iterations of attempting to estimate model parameters. When a model fails to converge, you as the researcher have several options, and this is a case where *researcher degrees of freedom*—the numerous seemingly innocuous choices made during the research process that enable researchers to find "statistically significant evidence consistent with any hypothesis" (Simmons et al., 2011)—might sneak in. As noted by Barr and colleagues (2013), simplifying the model should be done in a principled way, and you should not delete random slopes by default without first exploring whether these random effects may be necessary.

One way to address convergence issues is to add *control parameters* to your model. There are lots of control parameters, and depending on the source of the convergence issues, some may be more appropriate or useful than others. The one I recommend starting with is increasing the number of iterations before the model "gives up" on finding a solution. The model specification below is identical to the one above, with the exception that I have included a control parameter that increases the number of

[6] Response time data should really be analyzed with *generalized* linear mixed effects models (more on this in the section on binomial data) assuming an inverse Gaussian distribution and an identity link function because response times tend to be positively skewed (Lo & Andrews, 2015). For simplicity, however, we will use *general* linear mixed models via the `lmer()` function, which assumes a Gaussian distribution of response times; the parameter estimates change a bit when we assume an inverse Gaussian distribution, but the conclusions do not change. Mixed modeling is quite robust to violations of the normality assumption, so it is acceptable to use general mixed models here.

iterations and explicitly specifies the optimizer (i.e., the method by which the model finds an optimal solution):

```
> rt_full.mod <- lmer(RT ~ 1 + modality + (1 + modality | PID) + (1 + modality
| stim), data = rt_data, control = lmerControl(optimizer = “bobyqa”, optCtrl =
list(maxfun = 1e9)))
```

This model converges, great! Before we move on, another important thing to check for is *overfitting*. Even if your model converges, it is possible that your random effects structure is unnecessarily complex, and some of the random effects can (and arguably should; see Bates et al., 2015) be removed. To conceptualize why overfitting is a problem, imagine an analogous situation in ordinary least squares regression. If you are interested in predicting some outcome, you could theoretically include every predictor you can possibly come up with, and the model fit will only improve. This is because a fact of statistical modeling is that adding predictors can never remove explained variance—even if the predictor is seemingly unrelated to the outcome, adding it to the model can only increase variance explained (or leave it essentially unchanged). As an example, if you are interested in predicting response times to words, including a dichotomous variable indicating whether or not a person has eaten at least one bagel that week is unlikely to account for much unique variance in response times, so including it in your model would amount to overfitting. This is an extreme example, but the point is: just because a model converges does not mean all variables included are necessary, and a more parsimonious model might be preferred (Bates et al., 2015).

One way to check for overfitting is to examine the summary output for your selected model and look for correlations among random effects of -1.00 or 1.00. Perfect correlations are an indication that the model has been overfit, and perhaps one or more random effects are not accounting for significant unique variance in the outcome and should be removed. Here is the section of the summary output that is relevant to the issue of overfitting:

```
> summary(rt_full.mod)

Random effects:
 Groups   Name                  Variance   Std.Dev.   Corr
 stim     (Intercept)              12893     113.55
          modality                  1945      44.10    0.35
 PID      (Intercept)              51882     227.78
          modality                  9397      96.94   -0.12
 Residual                          98903     314.49
---
```

The column called `Corr` indicates that the correlation between the random intercept for stimulus and the by-stimulus random slope for modality is 0.35, and the correlation between the random intercept for participant and the by-participant random slope for modality is -0.12. Given that neither of these correlations is 1.00 or -1.00, and the contribution of each of the random effects is reasonably large (see the `Variance` column), overfitting does not appear to be an issue for this model (see the section on analyzing binomial data below for an example of a case where a model has been overfit). This output also contains important information about the random effects estimates. The two rows corresponding to the item random effects indicate the extent to which response times elicited by particular stimuli vary around the fixed intercept and slope. For example, the standard deviation for the by-item random intercept is 113.55, which indicates that item-level response times vary around the group estimate of 1,391 (see below) by about 114 ms on average. Similarly, the two rows corresponding to the participant random effects indicate the extent to which response times elicited by particular participants vary around the fixed intercept and slope. For example, the by-participant random slope is 96.94, which indicates that participant-level slopes vary around the estimated group slope of 96.39 (see below) by about 97 ms on average. Thus, an individual whose slope is one standard deviation below the mean will have an estimated slope near 0 (indicating that their response times are not affected by the modality in which the words were presented), whereas an individual whose slope is one standard deviation above the mean will have a very steep slope (indicating a difference between modalities of almost 200 ms).

Now that we have our model, how do we determine if modality actually affected response times? One method is to look at the summary output for the model, which involves running a Wald test. Note that the *lme4* package only provides *p*-values for model parameters from generalized linear mixed effects models, which are built using the `glmer()` function and will be described in the next section. If you are using the `lmer()` function, as we are in this example, and want to obtain *p*-values for model parameters, you will need to install and load another package, called *lmerTest* (Kuznetsova et al., 2017).[7] Here is the code and a section of the output that contains relevant information about the fixed effects:

```
> library(lmerTest)
> summary(rt_full.mod)

Fixed effects:
                   Estimate  Std. Error     df   t value  Pr(>|t|)
(Intercept)         1391.23       29.52  69.42     47.12    <2e-16
modality              96.39       14.88  64.01      6.48  1.51e-08
---
```

If you use this method, the *p*-value for the effect of modality appears in the column `Pr(>|t|)`. Recall that we used a dummy coding scheme such that the audio-only condition is the reference level, which will affect the interpretation of the fixed effects. The estimate for the fixed intercept indicates that in the condition coded 0 (audio-only), estimated response times are on average 1,391 ms across participants and items. The estimate for the fixed slope for modality indicates that response times in the audiovisual condition are on average an estimated 96 ms slower than those in the audio-only condition, again collapsing across participants and items.

Using the *p*-value from the summary output (accessed via the *lmerTest* package) is usually a fine approach, but a slightly preferable approach is to compare a model with the effect of interest (e.g., modality) to a model lacking that effect using a *likelihood ratio test* (see Detail Box 1). These tests are used to compare two *nested* models (the larger model contains all

[7] The *lmer* function does not include *p*-values because the null distribution is unknown (the error structure in multilevel models is complex, and the degrees of freedom cannot be calculated). Douglas Bates, one of the creators of the *lme4* package and the person who wrote the *lmer* function, has posted a helpful description of why he did not include *p*-values in that function (see Bates, 2006).

parameters in the reduced model, plus at least one additional parameter) by calculating the likelihood of the data under each of the two models—using a technique called maximum likelihood estimation—and statistically comparing those likelihoods (see Detail Box 2). If you obtain a small *p*-value from the likelihood ratio test, you reject the null hypothesis that the coefficients for the parameters that were removed to create the reduced model are zero, suggesting that the full model provides a better fit for the data.

Detail Box 1

The likelihood ratio test and the Wald test are asymptotically equivalent, meaning they give the same results as the sample size gets infinitely large. If you can get nested models to converge, you should stick with a likelihood ratio test, but the Wald test will almost always provide the same result (the *p*-value will be slightly different, but the conclusions will rarely change). One benefit of the likelihood ratio test is that if you have a fixed effect with three or more levels, the likelihood ratio test will serve as a significance test for the overall effect, but the Wald test will only provide the significance of each pairwise comparison to the reference level. The likelihood ratio test also allows you to simultaneously test the influence of multiple parameters at once (e.g., you can test whether a model including five fixed effects provides a better fit for the data than a nested model including only two of those fixed effects). See Engle for more details on the distinction between the likelihood ratio test and the Wald test (1984).

Detail Box 2

If you are using likelihood ratio tests to compare models differing in fixed effects, those models should be built using maximum likelihood (ML) estimation rather than *restricted* maximum likelihood (REML) estimation; this is because the REML estimate depends on the fixed components of the model, so comparison across models differing in their fixed components using this method is inappropriate (see Snijders & Bosker, 2012). To implement this in R, simply include `REML = FALSE` in the model specification. In contrast, REML should be used when comparing models differing only in random effects (variance estimates for the random effects are more precise with REML than with ML). To implement this in R, include `REML = TRUE` in the model specification, and `refit = FALSE` when you run the likelihood ratio test via the `anova()` command. Note that when you run a likelihood ratio test, the default is for *lme4* to automatically refit the models using ML; this default is in place to make it really difficult to test fixed effects using REML (which you should never do). However, if you are testing random effects, you can override the default with `refit = FALSE`. In our example, given that we are interested in whether modality (a fixed effect) influences response times, the models must be fit using ML rather than REML.

When we run a likelihood ratio test, we are basically asking “Does a model that includes information about the modality in which words are presented fit the data better than a model that does not include that information?” Here is how you perform this test in R (the *p*-value is in the `Pr(>Chisq)` column; see Detail Box 3):

```
> rt_reduced.mod <- lmer(RT ~ 1 + (1 + modality | stim) + (1 + modality |
PID), data = rt_data, control = lmerControl(optimizer = “bobyqa”, optCtrl =
list(maxfun = 1e9)))
> anova(rt_reduced.mod, rt_full.mod)

Data: rt_data
Models:
rt_reduced.mod: RT ~ 1 + (1 + modality | stim) + (1 + modality | PID)
rt_full.mod: RT ~ 1 + modality + (1 + modality | stim) + (1 + modality | PID)
                   Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
rt_reduced.mod      8 87896 87950  -43940    87880
rt_full.mod         9 87865 87926  -43924    87847 32.479      1  1.205e-08
---
```

Detail Box 3

The reason the column containing the *p*-value is called `Pr(>Chisq)` is because the test statistic for the likelihood ratio test—the difference in the *deviances* (defined as -2 times the log-likelihood) of the two models—is asymptotically χ^2 distributed with degrees of freedom equal to the difference in the number of parameters between the full and reduced models. In this example we have just one degree of freedom (which can be seen in the `Chi Df` column) because the only difference between the two models is the presence of the fixed effect for modality. The value in the `Chisq` column is the χ^2 value, so `Pr(>Chisq)` indicates the probability of observing a χ^2 value as extreme or more extreme than the value you observed (here, 32.48), assuming the null hypothesis is true. This is the definition of a *p*-value! In this case, the tiny *p*-value indicates that if the null hypothesis is true (i.e., there is no effect of modality on response times), the probability of observing data as extreme as what we observed is very low. We therefore reject the null hypothesis and assume that modality has an effect on response times.

In this example, our model converged when we increased the number of iterations (note that these models are computationally complex, so they can take a while to run). This will not always be the case, so I will point out a few other ways to address convergence issues. One option is to force the correlations among random effects to be zero. If you do not specify that the correlation is zero, the model attempts to estimate the correlation between these effects (for example, the correlation of -0.12 we observed above suggests that participants with larger intercepts are less affected by modality). If you are willing to accept that the correlation may be zero,[8] this will reduce the computational complexity of the model, and may allow the model to converge on parameter estimates. Note, however, that it is advisable to conduct likelihood ratio tests on nested models differing in the presence of the correlation parameter—or examine the confidence interval around the correlation—to determine whether elimination is warranted. If the model including the correlation parameter provides a better fit for the data than one without it or the confidence interval does not include zero, the correlation parameter should not be eliminated from the model. To remove this correlation in R, simply put a 0 where the 1 was in the random effects specification (here I did this for participants but not items). When you do this, however, it removes the random intercept, so you need to be sure to put it back into the model specification. Here is what the code would look like if you wanted to remove the correlation between the random intercept for participants and the by-participant random slope for modality:

```
> rt_full.mod <- lmer(RT ~ 1 + modality + (1 + modality | stim) + (0 +
modality | PID) + (1 | PID), data = rt_data)
```

Other options include changing the optimizer (e.g., `optimizer = “Nelder_Mead”`) or removing some of the derivative calculations that occur after the model has reached a solution using the following control parameter: `control = lmerControl(calc.derivs = FALSE)`. Finally, it may be that a model fails to converge simply because the random effects structure is too complex (Bates et al., 2015). In this case, one can selectively remove random effects based on their significance or contribution

[8] A situation in which you may not be willing to assume the correlation is zero is if that correlation is a crucial part of your research question. For example, if your research question addresses whether people with slower overall response times tend to be less affected by modality, then it would be critical to allow the model to estimate the correlation between the random effects. However, a typical study in experimental psychology is more interested in the fixed effects parameter estimates, so assuming the correlation is zero is often acceptable. The code to examine the confidence interval around standard deviation estimates is: `confint(rt_full.mod, parm = “theta_”, oldNames = F)`. The `parm` parameter indicates which parameters in the model will be given confidence intervals, and setting the `oldNames` parameter to `FALSE` simply gives the output more interpretable names.

to the total variance (i.e., remove those with low variance estimates), but it is important to consider the consequences of this decision. That is, if there is reason to expect that a particular random slope is important to include in the random effects structure, removing it may be ill-advised.

Now that we have our results, how do we present them in a manuscript? First, it is important to note that there are no explicit rules for reporting findings from model comparisons and the associated parameter estimates from the preferred model (Meteyard & Davies, 2019). How results are reported depends on the number and nature of model comparisons, the journal submission guidelines, and author and reviewer preferences. One way of reporting results is to report the χ^2 value from the likelihood ratio test and its associated *p*-value, as well as the beta coefficient, *t*-value, standard error, and *p*-value associated with the parameter of interest from the selected model. To report the findings described in the example above, a researcher could write: "A likelihood ratio test indicated that the model including modality provided a better fit for the data than a model without it ($\chi^2_1 = 32.48$; $p < .001$). Examination of the summary output for the full model indicated that response times were on average an estimated 96 ms slower in the audiovisual condition ($\beta = 96.39$, $SE = 14.88$, $t = 6.48$, $p < 0.001$)." However, I have had a number of reviewers tell me that this is too much information to include in the paper (a viewpoint I disagree with), and I have had others tell me that they appreciate the level of detail in the reporting of results. As long as you report your results transparently, the particular convention you follow is up to you (and of course making your data and code publicly available reduces the impact of this decision).

Analyzing data with a binary outcome (identification accuracy)

Now you should have a general understanding about how to test for a main effect of modality on response times. But what if you wanted to test for an effect of modality on accuracy at identifying words, where accuracy for each trial is scored as 0 or 1? These values are discrete and bounded by 0 and 1, so you need to use *generalized* linear mixed effects models; if we instead modeled this discrete outcome assuming a Gaussian distribution, the model would generate impossible predictions (e.g., a predicted probability of -0.2 or 1.3). The code is almost exactly the same as above, except rather than using the `lmer()` function you use the `glmer()` (**g**eneralized **l**inear **m**ixed **e**ffects **r**egression) function, and you need to include at least one additional parameter within the `glmer()` function indicating the assumed distribution of the dependent variable. You also may need to indicate a link function, which transforms the outcome into a continuous and unbounded scale. In this case, the discrete outcomes of 0 and 1 follow a *binomial* distribution, which should be modeled with logistic regression, typically using a *logit link function* (though other link functions, like the *probit*, are certainly possible). The logit link function transforms probabilities, which are bounded by 0 and 1, into a continuous, unbounded scale (log-odds). Using the logit link function allows us to model the linear relationship between the predictors and the log-odds of the outcome (which can be transformed back into odds and probabilities for ease of interpretation) without generating nonsensical predictions.

Put simply, the logit link function first transforms probabilities, which are bounded by 0 and 1, into odds, which are bounded by 0 and infinity—a probability of 0 corresponds to odds of 0, and a probability of 1 corresponds to odds of infinity. However, this scale still has a lower bound of 0, so the link function takes the natural logarithm (the logarithm of 0 is negative infinity, extending the lower bound of the scale from 0 to negative infinity) of the odds, resulting in the continuous and unbounded log-odds scale. This means that any predictions generated from the model will also be on a log-odds scale, which is not particularly informative, but luckily these predictions can be exponentiated to put them back on an odds scale, which in turn can be converted into probabilities.

Here is the code to build the full model:

```
> acc_full.mod <- glmer(acc ~ 1 + modality + (1 + modality | PID) + (1 +
modality | stim), data = acc_data, family = binomial)
```

This code is very similar to that for the response time analysis, with a few key differences. First, the dependent variable is `acc` (0 for incorrect and 1 for correct word identification) rather than `RT`. Since this outcome is binomially distributed, we indicate that we are using generalized linear mixed effects modeling by using the `glmer()` function, and we indicate that our dependent variable follows a binomial distribution with the additional parameter `family = binomial`.[9]

This model converged, but as described above it is important to check for overfitting. Below is the output of the `summary(acc_full.mod)` command. Note from the `Corr` column that the correlation between the random intercept for stimulus and the by-stimulus random slope for modality is 1.00, which is indicative of overfitting.[10]

```
summary(acc_model_full)

Random effects:
 Groups Name              Variance   Std.Dev.  Corr
 stim   (Intercept)       1.030e+00  1.015128
        modality          1.902e-05  0.004361  1.00
 PID    (Intercept)       4.149e-01  0.644127
        modality          4.933e-02  0.222098  -0.33
---
```

Given that this correlation is 1.00, and the by-stimulus random slope for modality appears to be contributing little to the total variance (see the value in the second row of the `Variance` column), we will remove that random effect and rebuild the full model:

```
> acc_full.mod <- glmer(acc ~ 1 + modality + (1 + modality | PID) + (1 |
stim), data = acc_data, family = binomial)
```

This model converged, and another look at the summary output indicates that this solved the overfitting problem. Further, a likelihood ratio test indicates that the maximal model did not provide a better fit for the data than the model without the by-stimulus random slope (see the R script), so we will stick with the model without the random slope. Next we will build a reduced model so that we can

[9] A logit link function is the default for logistic regression in R, so you do not need to include it in the model specification. If you wanted to change the link function to a probit link, for example, you would need to include `family = binomial(link = "probit")`.

[10] If you are following along in the R script, you probably noticed a warning message indicating that this model produced a singular fit. This is also an indication that the model is overly complex (which is typically attributable to an overly complex random effects structure). If this happens, you should check for correlations of 1.00 and -1.00, just as we did here, and remove random effects accordingly.

compare it to the full model to test for an effect of modality on word identification accuracy. This model is identical to the full model in all respects except that it lacks modality as a fixed effect:

```
> acc_reduced.mod <- glmer(acc ~ 1 + (1 + modality | PID) + (1 | stim), data =
acc_data, family = binomial)
```

It is important to note that although both the full and reduced models converged with this random effects structure and no control parameters, it is certainly possible (and indeed not uncommon) for the full model to converge but the reduced model to have convergence issues. In this case, you should find a random effects structure and combination of control parameters that enables both the full and reduced models to converge, because the models being compared via a likelihood ratio test should be nested and built with the same control parameters. That is, the models should be identical except for the presence of the fixed effect of interest. Here is the code and output for the likelihood ratio test assessing the effect of modality on word identification accuracy.

```
> anova(acc_reduced.mod, acc_full.mod)
Data: acc_data
Models:
acc_reduced.mod: acc ~ 1 + (1 + modality | PID) + (1 | stim)
acc_full.mod: acc ~ 1 + modality + (1 + modality | PID) + (1 | stim)
                       Df   AIC   BIC  logLik deviance  Chisq ChiDf Pr(>Chisq)
acc_reduced.mod         5 14695 14732 -7342.5    14685
acc_full.mod            6 14688 14732 -7337.7    14676 9.5901     1   0.001956
---
```

The small *p*-value associated with the likelihood ratio test indicates that the full model provides a better fit for the data than the reduced model, indicating that modality has a significant effect on spoken word identification accuracy.

Conclusions

Mixed effects modeling is becoming an increasingly popular method of analyzing data from experiments in which each participant responds to multiple items—and for good reason. The beauty of mixed effects models is that they can simultaneously model participant and item variability while being far more flexible and powerful than other commonly used statistical techniques: they allow for variable trial spacing across participants, they handle missing data well, they can seamlessly include continuous predictors, and they provide estimates for average (as well as by-participant and by-item) effects of predictors on the outcome.

However, as the saying goes: with great power comes great responsibility (Lee, 1962). These models can be easily implemented in R without cost, but it is important that researchers ensure that this powerful tool is used correctly. Indeed, although more and more researchers are implementing mixed effects models, there is a concerning lack of standard in how the models are both implemented and reported (Meteyard & Davies, 2020). Many analytical decisions must be made when using this statistical technique—consider, for example, the number of options available to the researcher if a model fails to converge—resulting in a massive number of "forking paths" (Gelman & Loken, 2014) that a research may embark upon to obtain statistically significant results. Given the considerable number of choices a researcher may make during data analysis (i.e., researcher degrees of freedom; Simmons et al., 2011), it is important that these models be used carefully and reported transparently (see Meteyard & Davies, 2020 for an example of how models and results should be reported)

The goal of this article was to serve as an accessible, broad overview of mixed effects models for researchers with minimal experience with this type of modeling, focusing on what they are, what they offer over other analytical techniques, and how to implement them in R. For more in-depth descriptions of mixed effects modeling, I recommend: Barr (2008) for analyzing eye tracking data; Barr et al. (2013) for an argument in favor of utilizing the maximal random effects structure justified by the design; Bates et al. (2015) for an argument in favor of using parsimonious mixed models; Baayen, Davidson, and Bates (2008) and Quené and van den Bergh (2008) for descriptions of mixed models with crossed random effects for participants and items; Judd et al. (2017) as well as Westfall et al. (2014) for an overview of design types and statistical power; and Jaeger (2008) for a description of logit mixed models. For a very accessible tutorial, consult Bodo Winter's tutorials in the "Helpful Links" section below (if you have limited statistical experience, I recommend starting with these). If you are interested in extending mixed effects modeling to growth curve analysis (which is used to model change over time, for example, changes in pupillary responses or fixation locations), see Mirman (2014) and Mirman (2008).

Helpful Links

Install R (Mac): https://cran.r-project.org/bin/macosx/
Install R (Windows): https://cran.r-project.org/bin/windows/base/
Install RStudio: https://www.rstudio.com/products/rstudio/download/
Bodo Winter's tutorial (part 1, linear models):
http://www.bodowinter.com/tutorial/bw_LME_tutorial1.pdf
Bodo Winter's tutorial (part 2, linear mixed effects models):
http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf

Acknowledgments: The author is grateful to Michael Strube for providing detailed feedback on an earlier draft of the paper, and to Julia Strand and Kristin Van Engen for providing helpful comments and suggestions throughout the writing process.

Funding: This work was supported by the National Science Foundation through a Graduate Research Fellowship awarded to Violet Brown (DGE-1745038).

Author contribution: V. A. Brown is fully responsible for the contents of this article. She devised the idea for the article, wrote the article in its entirety, wrote the accompanying R script, and generated the dummy data on which the models are based.

Conflicts of interest: The author declares that there were no conflicts of interest with respect to authorship or publication of this article.

Data, materials, and online resources: The dummy data and R script used to generate the models described in this paper are available at https://osf.io/v6qag/.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Barr, D. J. (2008). Analyzing “visual world” eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474.

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). https://doi.org/10.1016/j.jml.2012.11.001

Bates, D. (2006, May 19). *[R] lmer, p-values, and all that*. https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. In *arXiv [stat.ME]*. arXiv. http://arxiv.org/abs/1506.04967

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., & Green, P. (2014). *Package “lme4”* (Version 1.1-15). R foundation for statistical computing, Vienna, 12. https://github.com/lme4/lme4/

Enders, C. K. (2010). *Applied Missing Data Analysis (Methodology in the Social Sciences)* (1 edition). The Guilford Press.

Engle, R. F. (1984). Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics* (Vol. 2, pp. 775–826). Elsevier.

Etz, A. (2018). Introduction to the Concept of Likelihood and Its Applications. *Advances in Methods and Practices in Psychological Science*, *1*(1), 60–69.

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis—a “garden of forking paths”—explains why many statistically significant comparisons don’t hold up. *American Scientist*, *102*(6), 460–466.

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.

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*, 601–625.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, Articles*, *82*(13), 1–26.

Lee, S. (1962). *Introducing Spider-Man*. Amazing Fantasy #15.

Liben-Nowell, D., Strand, J., Sharp, A., Wexler, T., & Woods, K. (2019). The danger of testing by selecting controlled subsets, with applications to spoken-word recognition. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.51

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171.

Meteyard, L., & Davies, R. (2019). *Best practice guidance for linear mixed-effects models in psychological science*. https://doi.org/10.31234/osf.io/h3duq

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R (Chapman & Hall/CRC The R Series)* (1 edition). Chapman and Hall/CRC.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425.

Raudenbush, S. W. (1988). Educational Applications of Hierarchical Linear Models: A Review. *Journal* *of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *13*(2), 85–116.

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

Royston, P., Altman, D. G., & Sauerbrei, W. (2005). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*, 127–141.

Schober, P., & Vetter, T. R. (2018). Repeated Measures Designs and Analysis of Longitudinal Data: If at First You Do Not Succeed-Try, Try Again. *Anesthesia and Analgesia*, *127*(2), 569–575.

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.

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

Wendorf, C. A. (2004). Primer on Multiple Regression Coding: Common Forms and the Additional Case of Repeated Contrasts. *Understanding Statistics*, *3*(1), 47–57.

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.