tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Psychological Methods

1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000159

Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables that Vary Within-Subjects and/or Within-Items

AQ: 1

AQ: au

Markus Brauer and John J. Curtin
University of Wisconsin-Madison

Abstract

In this article we address a number of important issues that arise in the analysis of nonindependent data. Such data are common in studies in which predictors vary within "units" (e.g., within-subjects, within-classrooms). Most researchers analyze categorical within-unit predictors with repeated-measures ANOVAs, but continuous within-unit predictors with linear mixed-effects models (LMEMs). We show that both types of predictor variables can be analyzed within the LMEM framework. We discuss designs with multiple sources of nonindependence, for example, studies in which the same subjects rate the same set of items or in which students nested in classrooms provide multiple answers. We provide clear guidelines about the types of random effects that should be included in the analysis of such designs. We also present a number of corrective steps that researchers can take when convergence fails in LMEM models with too many parameters. We end with a brief discussion on the trade-off between power and generalizability in designs with "within-unit" predictors.

Translational Abstract

Researchers and practitioners sometimes want to analyze data that are "nonindependent." Data are said to be nonindependent when the study is designed such that certain data points can be expected to be on average more similar to each other than other data points. This is usually the case when each subject provides multiple data points (so-called within-subject designs), when subjects belonging to higher-order units influence each other (e.g., students clustered in classrooms, employees clustered in teams), or when subjects react to or evaluate the same set of items (e.g., pictures, words, sentences, products, art works, target individuals). In the present article, we propose that all types of nonindependent data can be analyzed with the same statistical technique called "linear mixed-effects models." Compared to standard statistical tests belonging to the family of "General Linear Models" (e.g., ANOVA, regression), linear mixed-effects models have a "complicated error term," i.e., the data analyst has to explicitly include all possible reasons for why the predictions of the statistical model may be wrong (these possible reasons are called "random effects"). It is not always obvious how to identify all possible sources of error. In this article, we provide clear guidelines on the type of random effects that researchers and practitioners should include when estimating linear mixed-effects models. Failure to include the appropriate random effects leads to an unacceptable false positive rate (or "type I error rate"), i.e., a high proportion of statistically significant results for effects that do not exist in reality.

AQ: 2

Keywords: The analysis of nonindependent data, within-subjects designs, linear mixed-effects models, fixed and random effects, convergence problems

AQ: 3

In recent years, interest in "linear mixed-effects models" (LMEMs) has increased drastically. Influential articles by Judd, Westfall, and Kenny (2012) and Barr, Levy, Scheepers, and Tily (2013) have made clear that many psychological studies require these types of models. The traditional ANOVA/regression approach is limited in that it poorly handles missing data and cannot handle continuous predictors that vary within "units" (e.g., within-subjects, within-groups, within-classrooms). More importantly, this approach yields biased inferential statistics when the same subjects are exposed to the same set of items (or stimuli or targets).

Markus Brauer and John J. Curtin, Department of Psychology, University of Wisconsin-Madison.

The authors thank Joe Austerweil, Harald Baayen, Mitchell Campbell, Marna Dunne, Patrick Forscher, Andrew Gelman, Charles Judd, Hannes Matuschek, Serban Musca, Kristopher Preacher, Michael Ro, Tom Snijders, Adrienne Wood, and Martin Zettersten for feedback on earlier drafts of this article. We are also grateful to Dale Barr, Douglas Bates, Ben Bolker, Joop Hox, and Jonathan Templin who answered questions by e-mail. Special thanks go to Jake Westfall who helped us enormously by always being available to discuss data-analytic questions with us at a moment's notice and sometimes at great length. No information in this article was disseminated prior to its publication.

Correspondence concerning this article should be addressed to Markus Brauer, Department of Psychology, University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706-1611. E-mail: markus.brauer@wisc.edu

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Only mixed models yield unbiased parameter estimates with acceptable type-I and type-II error rates. AQ: 4

The increased use of LMEMs will shape the types of studies that psychologists will conduct in the future. We now know that we should think about items the same way we have always thought about subjects: Ideally, both should be sampled from a larger pool of possible exemplars, and we typically want to generalize our findings from the sample to the entire population (Clark, 1973). In order to do so, it is necessary to randomly select a sufficiently large sample of subjects and a sufficiently large sample of items (Bahník & Vranka, 2017). Studies with only one observation per level of the within-subject design suffer from limited generalizability. The same is true for other elements of our studies: We probably will want to generalize our findings beyond the specific groups, classrooms, confederates, tasks, and locations that we included in our study. As such, the role of LMEMs in data analysis will likely increase in the future: Most published studies will require LMEMs, and a solid mastery of LMEMs will be crucial for designing and analyzing impactful research.

Although many researchers have recognized the need to analyze their data with LMEMs, not all of them know how to correctly specify these models for a variety of designs. The purpose of this article is to provide clear guidance on the analysis of data with one or more sources of nonindependence. Specifically, we will describe, in a user-friendly and pragmatic way (a) the type of effects that should be included in models examining data from different designs, and (b) how to address a variety of problems that one might run into when estimating LMEMs.

The article addresses itself to a variety of audiences. The initial sections target LMEM novices. We start out with an introduction to linear mixed-effects models and the analysis of dichotomous predictor variables that vary within units. We also define fixed effects and random effects. Advanced LMEM beginners may want to skip to the section on the analysis of continuous within-subjects predictors. We show that dichotomous and continuous within-subjects predictors can be analyzed using the same conceptual framework and with virtually identical commands in most data analysis programs. We extend the framework to multilevel models. The experienced LMEM user may be most interested in the section on "Multiple Sources of Nonindependence" and the remainder of the article. These sections contain guidance on the types of random effects that should be included depending on whether each predictor in the model varies either within-units or between-units (e.g., subjects, items, classrooms). We also propose corrective steps that researchers can take to simplify their models when they run into convergence problems. The article ends with a brief discussion on statistical power and generalizability.

Terminology and Data Formats

In this article, we will use the term "linear mixed-effects models" (LMEMs) to refer to models with one or more random effects. These models include data analytic techniques like hierarchical regression, hierarchical linear modeling (HLM), multilevel regression, multilevel linear modeling, linear mixed models, and random coefficient models. The common characteristic of these models is that they allow researchers to analyze data with one or more sources of nonindependence. Data are nonindependent when multiple data points are collected from each subject (e.g., within-subject design, longitudinal research) or when subjects belong to groups and group members influence each other (e.g., subjects belong to the same family or discussion group, students are nested in classrooms). The data are also nonindependent when subjects are exposed to the same set of items (e.g., subjects react to pictures, judge words, or evaluate target individuals). Among the five assumptions of ANOVA and regression—exact X, independence, normality, constant variance, and linearity—a violation of the independence assumption is generally considered the most serious one in that it produces the most incorrect inferential statistics (Judd, McClelland, & Ryan, 2009). In this article, we will frequently use examples in which the predictor variable varies within-subjects, but readers should be aware that the presented data-analytic techniques can also be used when the predictor varies within higher-order "units," such as groups or classrooms.

In order to adopt the "LMEM way of thinking," an increasing number of researchers now analyze their data in "long format" rather than "wide format" when there are multiple observations per subject. When data are presented in wide format, there is one row per subject and the multiple observations appear in different columns. In long format, there is one row per observation and thus multiple lines per subject. Imagine a hypothetical experiment (which we will refer to as "Study 1" throughout the article) in which researchers recruit 100 undergraduates and ask them to list two high-prestige classes and two low-prestige classes they took in college. Prestige is defined as the extent to which a class "looks good" on the students' transcript (e.g., difficult science classes, honors classes, graduate level classes, classes that help students get a job or be admitted to graduate school). Students are then asked to indicate, on a 9-point scale, their liking for each of the four classes. Let's further assume that the study takes place at a large public university so that each student evaluates a different set of four classes. The experiment has a single dichotomous within-subjects predictor, which we will call "prestige." The outcome variable is "liking." The data in wide format would contain 100 rows (and, among other variables, four columns corresponding to the liking for each of the four classes), whereas the data in long format would contain 400 rows (and all four liking ratings from the same subject would appear in the same column; see Table 1 for an example.). T1

LMEMs require data to be in long format. In addition, many researchers argue that data files in long format are more "tidy" in that they resemble those of between-subjects designs: There is one column that represents the dependent variable and one or more columns that correspond to the predictor variable(s), and one observation is associated with one row (Wickham, 2014). The `reshape` function in R allows researchers to transform data files from wide format into long format, and vice versa (see Appendix A).

Although the data format has no impact on the conclusions—both formats yield statistically identical results when analyzed with the appropriate models—a long data format helps researchers understand the underlying logic of LMEM, which can then easily be generalized to more complex designs, such as studies with a continuous within-unit predictors or studies in which there are multiple sources of nonindependence. Data files in long format, however, require a different syntax in most data analysis programs.

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Table 1
The Same Data in Wide Format (Top) and Long Format (Bottom)

Row	subject.ID	like.lo1	like.lo2	like.hi1	like.hi2	gender	age
1	1	6	5	7	8	1	20
2	2	1	4	3	3	2	18
3	3	7	6	7	5	2	21
4	4	7	8	9	9	1	19
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
100	100	4	5	7	5	2	20

Row	subject.ID	like	prestige	gender	age	item.ID
1	1	6	1	1	20	1
2	1	5	1	1	20	2
3	1	7	2	1	20	3
4	1	8	2	1	20	4
5	2	1	1	2	18	5
6	2	4	1	2	18	6
7	2	3	2	2	18	7
8	2	3	2	2	18	8
9	3	7	1	2	21	9
10	3	6	1	2	21	10
11	3	7	2	2	21	11
12	3	5	2	2	21	12
13	4	7	1	1	19	13
14	4	8	1	1	19	14
15	4	9	2	1	19	15
16	4	9	2	1	19	16
. . .	. . .	. . .	. . .	. . .	. . .	. . .
400	100	5	2	2	20	400

Note. Only the data from the first four subjects and the last row in the data file are shown. Low-prestige classes are referred to with "lo" (top panel) or "1" (bottom panel), whereas high-prestige classes are referred to with "hi" (top panel) or "2" (bottom Panel). Although not necessary, we added a variable called "item.id" to the data file in long format to make explicit that each subject evaluated his/her own set of four classes.

As we will see, this syntax is often shorter and more intuitive than the one we use when data are in wide format.[1] Fn1

Introduction to Linear Mixed-Effects Models

Consider the hypothetical experiment presented above (Study 1): One hundred subjects rate the extent to which they liked two high-prestige classes and two low-prestige classes. Table 2 presents the R script for the data in the traditional wide format. If the difference score is statistically different from zero, then subjects' liking for the high-prestige classes is reliably different from that for the low-prestige classes. T2

Table 2
R Script for Hypothetical Study 1, When the Data are in Wide Format

```
d <- dfReadDat ("data_Study1_wide.dat")
d$ave_like_lo <- (d$like_lo1 + d$like_lo2)/2
d$ave_like_hi <- (d$like_hi1 + d$like_hi2)/2
d$difference <- d$ave_like_hi − d$ave_like_lo
model_1a <- lm(difference ~ 1, data = d)
summary(model_1a)
```

Note. The Study contains a single dichotomous predictor variable that varies within-subjects (prestige). The term "lm" stands for "linear model". Here the difference score (liking for high-prestige classes minus liking for low-prestige classes) is regressed on the intercept b_0 (labeled "1" in R). In other words, the model tests whether the difference scores are on average reliably different from zero.

Table 3 shows the R script when the data are in long format. As it turns out, the script is rather similar to the one we would have used if the predictor ("prestige") had varied between-subjects: An outcome variable is regressed on (the intercept and) a dichotomous predictor. The only difference is that the model statement now contains an additional element (`1 + prestigeC|subject.ID`). Expressed in a very simplified way, the additional element tells the data analysis software that the predictor "prestige" varies within-subjects. T3

As already mentioned, the two analyses reported in Table 2 and 3 yield identical results for the parameter estimates and their standard errors, the *df*s, the *F*- and *p* values.[2] For ease of interpretation we have recoded the dichotomous prestige variable into −.5 Fn2

[1] Although we will be providing only R script throughout this article, most of the described analyses can be performed with other major data analysis programs (e.g., SAS, SPSS).

[2] The two analyses yield identical results only if the LMEM uses the Kenward-Roger method to compute the degrees of freedom and no constraints are imposed on the covariance matrix of the LMEM (both are the default in R). We will present different methods to compute the degrees of freedom in LMEMs in the section on Restricted Maximum Likelihood. Throughout the article, we will present LMEMs with an "unstructured covariance matrix," that is, a covariance matrix upon which no constraints have been imposed.

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Table 3
R Script for Hypothetical Study 1 With Data in Long Format

```
library (lme4)
library (car)
d <- dfReadDat ("data_Study1_long.dat")
d$prestigeC <- d$prestige - 1.5
model_1b <- lmer(like ~ 1 + prestigeC + (1 +
  prestigeC|subject.ID), data = d)
summary(model_1b)
Anova(model_1b, type = 3, test = "F")
```

Note. The library statements load the packages needed to perform the analyses. The package "lme4" was written by Bates, Mächler, Bolker, and Walker (2015b). The package "car" was written by Fox and Weisberg (2011). The term "lmer" stands for "linear mixed-effects in R". The summary statement produces the parameter estimates, the ANOVA statement the inferential statistics.

and +.5 so that it is centered around zero. Such recoding is not necessary, as it will not affect the parameter estimate for the prestige effect. It simplifies, however, some explanations given below.

The LMEM approach in Table 3 helps us think about within-unit analyses in a different way than the one that is taught in many traditional statistics textbooks. As it turns out, there are numerous similarities and only a few differences between a purely between-subjects analysis and a within-subjects analysis that takes the form of a LMEM (assuming the only difference between the designs is whether the predictor of interest varies between-subjects or within-subjects). We will discuss the similarities and differences in turn.

Fixed Variables and Random Variables

In both types of analyses, the relevant variables are considered to be either "fixed" or "random" (Kreft & DeLeeuw, 1998). In the example above, `prestige` is a fixed variable (with two levels), whereas `subject.ID` is a random variable (with 100 levels). A variable is considered *fixed* when data have been gathered from all the levels of the variable that are of interest. It is also assumed that the values of a fixed variable in one study are the same as the values of the fixed variable in another study. The variables that are (or might be) implicated by theoretical predictions tend to be fixed, which is why predictor variables are nearly always fixed.

A variable is considered *random* when it has many possible levels and when the researchers' interest is in all possible levels, but only a random sample of levels is included in the data. Subjects that are randomly selected from a larger pool of possible subjects and items that are randomly selected from a larger pool of possible items are nearly always random variables. Other typical random variables are individuals who work with multiple subjects (e.g., managers, teachers, therapists, social workers), higher-order units that subjects are nested in (e.g., families, work teams, classrooms, counties), and settings (e.g., locations on campus or in town, different situations in which a behavior may occur).

The levels of random variables are usually nominal in nature, that is, the numbers assigned to them have no meaning except that they allow us to distinguish the different exemplars. A variable that assigns each "unit" (i.e., subject, item, manager, classroom, setting) a different identification number is usually random (see `subject.ID` and `item.ID` in Table 1), whereas variables that describe characteristics of these units are usually fixed. In Table 1 for example, the fixed variable `age` describes a characteristic of the units identified by the random variable `subject.ID`, and the fixed variable `prestige` describes a characteristic of the units identified by the random variable `item.ID`. This is why measured predictors, covariates, and demographics are generally fixed variables: They might be implicated by theoretical predictions, they describe characteristics of the subjects, and it generally assumed that the study includes a large enough sample so that data have been collected from all the levels of the variable that are of interest.

Random variables are explicitly included in the data analyses only if there is more than one observation per level of the variable. This is why `subject.ID` is not included as a predictor in the analyses of purely between-subjects designs (e.g., independent-samples *t* test, standard ANOVA, multiple regression). In Study 1, however, each subject made four ratings, two for the high-prestige classes and two for the low-prestige classes. As a consequence, the variable `subject.ID` has to be explicitly included in the data analyses. Said differently, random variables are included in the analyses only if they create nonindependence in the data. In Study 1, the four ratings from the same subject are clearly not independent from each other, and this is why `subject.ID` is part of the R script in Table 3.

Simple Versus Complex Error Terms

Statistical analyses all have the same basic structure: DATA = MODEL + ERROR (see Table 4). Every statistical model makes predictions based the (weighted) mean of the outcome variable (β_0) and one or more predictors (here: $\beta_1 X$). The major difference between the analyses of independent data (e.g., between-subjects analyses) and nonindependent data (e.g., within-subject analyses) is the complexity of the error term.

In the analyses of independent data, the error term is relatively simple: It only has one element, the random error. When data are nonindependent and analyzed via a LMEM, the error term usually consists of multiple components. This is because there are multiple reasons the model predictions may be incorrect in these models. One source of error, like in any model, is differences between subjects in general. For example, subjects in Study 1 may differ in how they use the rating scale or the extent to which they enjoy university classes in general. This source of error is often referred to as "random intercept" or, to make explicit that the source of error is caused by subjects, "by-subject random intercept" (labeled "e_{RI}" and "u_{0j}" in Table 4). A second source of error stems from differences between subjects in how they are affected by the predictor variable(s). In Study 1, subjects may differ in the extent to which they prefer high-prestige classes over low-prestige classes. This source of error is often referred to as "by-subject random slope" or simply "random slope" (labeled "e_{RS}" and "$u_{1j}X_{ij}$" in Table 4). A third source of error is random error (labeled "e" and "e_{ij}" in Table 4). Just like in the between-subjects case, this element captures

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Table 4
Comparison Between the Analyses of Independent and Nonindependent Data

Independent data (e.g., between-subjects designs)					Nonindependent data (e.g., within-subjects designs)					
DATA	=	MODEL	+	ERROR	DATA	=	MODEL	+	ERROR	
DATA	=	MODEL	+	SIMPLE ERROR TERM	DATA	=	MODEL	+	COMPLEX ERROR TERM	
Y	=	$\beta_0 + \beta_1 X$	+	e	Y	=	$\beta_0 + \beta_1 X$	+	$e_{RI} + e_{RS}$	+ e *
Y_i	=	$\beta_0 + \beta_1 X_i$	+	e_i	Y_{ij}	=	$\beta_0 + \beta_1 X_{ij}$	+	$u_{0j} + u_{1j} X_{ij}$	+ e_{ij} *
		Fixed effects		Random error			Fixed effects		Random effects	Random error

Note. The most important difference is the complexity of the error term. In both types of analyses, the hypothesis focuses on the fixed effects, most likely β_1. The table contains two equations per type of analysis, one simplified version without subscripts and one complete version (with subscripts) that can be found in numerous texts on LMEMs. If X is dichotomous, the model on the left is equivalent to an independent-samples *t*-test, Whereas the model on the right is equivalent to a paired-samples *t*-test (after averaging across multiple observations in the same cell of the within-subject design).
* $e_{RI} = u_{0j}$ = random intercept; $e_{RS} = u_{1j}X_{ij}$ = random slope.

all other sources of error, such as unreliable measurement and random fluctuations in ratings from one class to the next.[3]

ANOVA, *t* tests, and multiple regression are all special cases of the general linear model (GLM), which in turn is a special case of a LMEM: A GLM is a LMEM without random effects. Despite the differences in error terms, our focus in both GLMs and LMEMs is the interpretation of the regression coefficient associated with the predictor variable(s). In both columns of Table 4, β_1 represents the model's estimate of the effect of "prestige" on liking. If β_1 is statistically significant, we conclude that there is an effect of the predictor on the outcome variable.[4]

Interpretation of Fixed Effects and Random Effects

In both models in Table 4, the coefficients β_0 and β_1 are called "fixed effects." The interpretation of the fixed effects in a LMEM is straightforward, as it closely follows the interpretation of fixed effects in a standard GLM. The coefficients β_0 and β_1 test whether there is an effect of prestige on liking (β_1, the so-called "fixed slope") and whether subjects' predicted liking scores for classes with a score of zero on prestige (centered) are reliably different from zero (β_0, the so-called "fixed intercept"). These effects are called "fixed" because they apply to the entire sample. Like regular regression analysis, the test of the fixed intercept in a LMEM is conceptually relevant only if a score of zero on the predictor variable is a meaningful value (which often is only the case if the predictor has been centered around zero).

The random intercept (u_{0j}) and the random slope ($u_{1j}X_{ij}$) in the LMEM are called random effects because they represent the extent to which the coefficients β_0 and β_1 vary from one subject to the next. This point can easily be understood by rearranging the LMEM equation as follows:

$$Y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})X_{ij} + e_{ij} \quad (1)$$

This version of the equation illustrates that the LMEM estimates multiple components. Applied to Study 1, the first parenthesis ($\beta_0 + u_{0j}$) refers to the average like ratings, that is, the averages of the four like ratings per subject. The model estimates two entities, the mean of the average like ratings (β_0) and the extent to which subjects' average like ratings vary around this mean (u_{0j}). Although the model predicts one average like rating for the entire sample (the fixed intercept), it also allows for the 100 individual average like ratings to vary around this prediction. Each subject's average like rating will deviate to some extent from the fixed intercept. In other words, there are 100 u_0's in Study 1, one for each subject.

The second parenthesis in the equation ($\beta_1 + u_{1j}$) describes the effect of prestige on liking. The model estimates two entities, the mean of the preferences for one type of class over the other (β_1) and the extent to which subjects' preferences vary around this mean (u_{1j}). As with the average ratings, the model predicts one mean "prestige effect" for the entire sample (the fixed slope), but it also allows for the 100 individual preferences to vary around this prediction. Each subject's prestige effect (the extent to which s/he prefers high-prestige classes over low-prestige classes) will vary somewhat from the fixed slope, which is why there are 100 u_1's in Study 1.

A Linear Mixed-Effects Model Estimates Variances

It turns out that the LMEM does not estimate each of the 100 u_0's and each of the 100 u_1's. Instead, it estimates their variances. In other words, the computer estimates one parameter that repre-

[3] There are several other minor differences between the two types of models with regard to the notation that is used in most texts on data analysis. The variables and the error term in the between-subjects case have one subscript (i for subject), whereas they have two subscripts in the within-subjects case (usually j for subject and i for item). In our hypothetical Study 1, j varies between 1 and 100 (there are 100 subjects) and i varies between one and four (each subject rates four classes), so that there are in total 400 Y-values, 400 X-values, and 400 e-values.

[4] Some data analysts argue that ratings on Likert scales should be analyzed with ordered logit regression (and generalized linear mixed-effects models; see Fullerton & Xu, 2016) rather than repeated measures ANOVA or LMEMs. Strictly speaking, ratings on Likert scales are ordinal outcomes. To keep the explanations in this article simple, we are assuming throughout the article that outcomes are continuous.

sents the variance of the 100 individual average like ratings and one parameter that represents the variance of the 100 individual prestige effects. The model can thus account for the between-unit variability in the intercept or slopes without having a large number of parameters. It also estimates the variance of the random errors, the e_{ij}'s. This is why the random effects and the random error are sometimes referred to as "variance components" in a LMEM.[5] Fn5 Although the classic *F* test is usually framed in terms of Sum of Squared Errors (SSE), it can easily be shown that the variance of the residuals (the e_i's in the left panel of Table 4) is equivalent to the SSE divided by N-1. Thus, the classic *F* test can be seen as a comparison between models having different variance components.

Using the variable names from hypothetical Study 1, the equation for the LMEM is as follows (see also right panel of Table 4):

$$\text{like} = \beta_0 + \beta_1\text{prestigeC} + u_0 + u_1\text{prestigeC} + e \qquad (2)$$

By including a random intercept in the model (i.e., by estimating the variance of the u_{0j}'s), we are allowing for the possibility that subjects differ in their average liking for the four classes. It is theoretically possible to specify a model without a random intercept (which is equivalent to fixing the variance of the u_{0j}'s to zero). Such a model would make the assumption that subjects' predicted like ratings for classes with a score of zero on the centered prestige variable are all the same. This assumption is likely to be incorrect.

The same reasoning can be applied to the random slope: By including a random slope in the model (by estimating the variance of the u_{1j}'s), we are allowing for the possibility that subjects differ in the extent to which they prefer one type of class over the other. As with the random intercept, it is theoretically possible to fix the variance of the u_{1j}'s to zero (i.e., to estimate a model without a random slope). Such a model would make the (probably incorrect) assumption that subjects all have the same relative preference for high-prestige classes, that is, that the difference between liking for the high-prestige classes and liking for the low-prestige classes is the same for all subjects.

Remember that a repeated-measures ANOVA with the data in wide format and a LMEM with the data in long format yield identical results.[6] Fn6 This is because the repeated-measures ANOVA allowed subjects to vary in their average like scores and in their relative preference of one type of class over the other, just like the LMEM. In both models, a greater variability around the average effect translates into larger standard errors and larger *p* values. Note that including a random effect does not necessarily mean that a degree of freedom is used up. It simply means that certain entities are allowed to vary from one subject to the next.

Most importantly, the inferential statistics for a given fixed effect will maintain a type-I error rate of 5% only if the model also includes its corresponding random effect (Barr et al., 2013). Applied to Equation 2, this means that the test of β_1 will have an acceptable type-I error rate only if u_{1j} (the random slope) is included in the model, and the test of β_0 will have an acceptable type-I error rate only if u_{0j} (the random intercept) in included in the model. We will come back to this important point later.

Random Variables Versus Random Effects

It is essential to distinguish between random variables and random effects. Technically speaking, the data file in long format for our hypothetical Study 1 contains two random variables, `subject.ID` (with 100 levels) and `item.ID` (with 400 levels, see last column of bottom panel in Table 1). Each class (item) that is being evaluated in the study has its own identification number. Given that each student evaluates a different set of four classes, there are in total 400 classes (items) being evaluated in the study. Both subjects and items are random variables. Both have been selected from a larger pool of possible exemplars, and the researchers would like to generalize their results to all subjects (students) and all classes (items).

And yet, only one of these random variables, `subject.ID`, is explicitly included in the analyses: The LMEM includes two effects related to subjects, a by-subject random intercept and a by-subject random slope. However, the LMEM includes no effects related to items. This is because there is only one observation for each level of the (random) variable `item.ID`. Each class is being evaluated only once. This is an important take-home message: Not every random variable requires a random effect, and certain random variables may require more than one random effect. There is no one-to-one correspondence between random variables and random effects. In later sections of this article we will discuss what types of random effects, if any, should be included for each of the random variables in the data file.

Extension to Designs With Multiple Dichotomous Predictors

The LMEM approach described above can easily be extended to more complex designs. For example, imagine a 2 × 2 mixed-model ANOVA with one within-subject factor (e.g., prestige) and one between-subjects factor (e.g., gender). Like before, each subject rates four classes. It can be shown that this 2 × 2 mixed-model ANOVA is mathematically equivalent to a LMEM with one fixed intercept (β_0), one fixed effect for the within-subjects factor (β_1), one fixed effect for between-subjects factor (β_2), one fixed interaction effect (β_3), one by-subject random intercept (u_0), and one by-subject random slope for the within-subject factor (u_1). The equation and the R script for such a model are shown in Table 5. T5

Note that, like before, the "complex error term" contains three elements, between-subjects variation in the average ratings (u_0), between-subjects variation in how they are affected by the prestige manipulation (u_1), and random error. There is no between-subjects variation in how subjects are affected by gender, because each subject is either male or female. For the same reason, there is no

[5] The data analysis program also estimates the covariance between the random effects. We will come back to this point in the section on the number of parameters being estimated.

[6] Westfall has demonstrated that there are rare cases in which the two types of analyses yield slightly different results (http://stats.stackexchange.com/questions/117660/what-is-the-lme4lmer-equivalent-of-a-three-way-repeated-measures-anova). The reason for this is that for data sets with a negative intraclass correlation, the best-fitting repeated-measures model implies that some of the variance components underlying the data must be negative. But most mixed model programs have built-in constraints that require the variance component estimates to be nonnegative. So the mixed model will do the best that it can within its constraints, but it will never reach the repeated-measures ANOVA solution. According to our own simulations, such discrepancies are very rare and extremely minor in that they are visible only in the third or fourth decimal of the *F* value.

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

between-subjects variation in how subjects are affected by the gender by prestige interaction. Gender varies between subjects, and between-subjects variables cannot be an additional source of error.

Let's next consider a 2 × 2 fully within-subjects ANOVA. One might imagine a study in which students are asked to list eight classes, four highly prestigious ones and four less prestigious ones. Within each group, two are science classes and two are nonscience classes. The study has a 2 × 2 design with two within-subjects factors, prestige and science. There are two observations per cell of the design. The data from such a study can be analyzed with a repeated-measures ANOVA with data in wide format. An equivalent approach would be to enter the data in long format (eight lines per subject) and to specify a LMEM with one fixed intercept (β_0), one fixed effect for the first within-subjects factor (β_1), one fixed effect for the second within-subjects factor (β_2), one fixed interaction effect (β_3), one by-subject random intercept (u_{0j}), one by-subject random slope for the first within-subject factor (u_1), one by-subject random slope for the second within-subject factor (u_2) and one by-subject random slope for the interaction (u_3). The equation and the R script for such a model can be seen in Table 6.

The "complex error term" now contains five elements because there are five different reasons for why the model predictions may be "off:" the by-subject random intercept (u_0; to account for differences in scale usage and general liking of university classes), three by-subject random slopes (u_1, u_2, and u_3; to account for variation in how subjects are affected by the two within-subject factors and their interaction), and the random error (e).

More complex designs—for example, a 2 × 2 × 2 × 2 ANOVA with two between- and two within-subjects factors, or a mixed model with one dichotomous within-subject predictor and one continuous between-subjects predictor—can easily be analyzed within the LMEM framework. It is easy to include continuous between-subjects predictors (e.g., a score on an individual difference measure). Given that between-subjects predictors require no by-subject random slope, such predictors can simply be added to the fixed part of the model. As we will see in the next section, the LMEM framework can also accommodate continuous predictors that vary within subjects, which cannot be appropriately handled by a repeated-measures ANOVA.

Like in ANOVA and regression analysis, LMEMs usually require us to center the predictors prior to the analysis when the model contains an interaction term. More precisely, the tests of the lower-order effects typically answer theoretically meaningful questions only if the dichotomous variables have been recoded into −.5 and +.5 (or any other two values centered around zero) and continuous variables have been "mean-centered" (by subtracting the mean from each score; see Schielzeth, 2010). Consider the model in Table 6. Given that "scienceC" is coded −.5 and +.5, β_1 tests the effect of prestige on liking for classes that are conceptually half-way between the science classes and nonscience classes, that is, the main effect of prestige in the 2 × 2 ANOVA. If the predictor had been coded 1 and 2 ("science"), then the coefficient β_1 would test the effect of prestige on liking for classes that have a score of 0 on science, a conceptually meaningless test.[8]

It is generally advised to include one random slope for each within-subjects predictor, in addition to the random intercept. Note that lower- and higher-order interactions among within-subject predictors are themselves considered within-subject predictors. Thus, a study with three within-subject predictors and their interactions requires one random effect for each of the eight fixed effects (one random intercept, three random slopes for the main effects of the predictors, three random slopes for the two-way interactions, and one random slope for the three-way interaction). However, a study with two between-subjects predictors and two within-subjects predictors has 16 fixed effects (the intercept and all possible two-, three-, and four-way interactions), but requires only four random effects: the random intercept, the random slope for the first within-subjects predictor, the random slope for the second within-subjects predictor, and the random slope for the interaction among the two within-subjects predictors.

Table 5
The LMEM and the R Script for a Study With One Dichotomous Within-Subject Variable ("Prestige") and One Dichotomous Between-Subject Variable ("Gender")

$$\text{like} = \beta_0 + \beta_1\text{prestigeC} + \beta_2\text{genderC} + \beta_3\text{prestigeC} * \text{genderC} + u_0 + u_1\text{prestigeC} + e \quad (3)$$

```
d$prestigeC <- d$prestige - 1.5
d$genderC <- d$gender - 1.5
model_1c <- lmer(like ~ 1 + prestigeC * genderC +
  (1 + prestigeC | subject.ID), data = d)
summary(model_1c)
Anova(model_1c, type = 3, test = "F")
```

Note. Both independent variables are being centered in order to be able to interpret the lower-order effects (the "main effects").[7]

Table 6
The LMEM and the R Script for a Study With Two Dichotomous Within-Subject Variables ("Prestige" and "Science")

$$\text{like} = \beta_0 + \beta_1\text{prestigeC} + \beta_2\text{scienceC} + \beta_3\text{prestigeC} * \text{scienceC} + u_0 + u_1\text{prestigeC} + u_2\text{scienceC} + u_3\text{prestigeC} * \text{scienceC} + e \quad (4)$$

```
d$prestigeC <- d$prestige - 1.5
d$scienceC <- d$science - 1.5
model_1d <- lmer(like ~ 1 + prestigeC * scienceC +
  (1 + prestigeC * scienceC|subject.ID), data = d)
summary(model_1d)
Anova(model_1d, type = 3, test = "F")
```

Note. Both independent variables are being centered in order to be able to interpret the lower-order effects (the "main effects").

[7] It is not necessary to specifically mention the fixed and random intercepts in the R script because R assumes that you want to include these intercepts. The following R script produces the same output as the R script presented in Table 5: `model_1c <- lmer(like ~ prestigeC * genderC + (prestigeC|subject.ID), data = d)`. For pedagogical purposes, we decided to specifically mention all intercepts in the R scripts presented in this article.

[8] This will be true in R only if the predictors are coded as numeric variables. Although R users have the possibility to code their variables as factors, this practice should be avoided when estimating LMEMs as lme4 has problems analyzing predictor variables that are coded as factors.

The Analysis of Continuous Within-Subject Predictors

The LMEM approach can easily be extended to the analyses of continuous within-unit variables (e.g., within-subjects, within-classrooms). To illustrate this type of analysis imagine a study (hypothetical Study 2) in which the experimenter asks 50 students to list eight classes they took in college and to rate both the perceived prestige of each class (the predictor variable; from 1 = *not prestigious at all* to 9 = *highly prestigious*) and their liking of it (the outcome variable; from 1 = *don't like at all* to 9 = *like a lot*; see Table 7 for an example data file). These data cannot be analyzed via a standard ANOVA or any other statistical procedure belonging to the general linear model, but require a LMEM (Hox, 2010).

The LMEM used to analyze the data from this study contains two fixed effects (the intercept and the effect of prestige), one random intercept, one random slope, and the residuals. The model equation and the R script to analyze the data are shown in Table 8. Note that the equation for the LMEM and the model statement in the R script are identical to those for Study 1 (see Equation 2 and Table 3). In the general linear model, dichotomous between-subjects predictors are analyzed the same way as continuous between-subjects predictors, in that they are entered as predictor variables in the equation. The same is true with LMEMs: continuous and dichotomous within- subject predictors are analyzed with same conceptual framework and using the same script to describe the model to be estimated.

Different Forms of Mean-Centering

Although the predictor variable is centered around zero in both Studies 1 and 2, there is an important difference between the two: In Study 1, the dichotomous predictor is recoded into −.5 and +.5, and this recoding is not necessary, because it will not affect the parameter estimate of the fixed effect associated with the predictor. In Study 2, however, the continuous predictor is centered around each subject's own mean (sometimes referred to as "cluster-mean centering" or "group-mean centering"). This manipulation is necessary to obtain an unbiased estimate of the within-subject association between the predictor and the outcome. Failure to center the continuous within-subject predictor or other types of mean-centering—for example, centering the predictor around its grand mean, that is, the mean of all 400 prestige ratings in Study 2—will produce estimates that are "uninterpretable" in most cases (Raudenbush & Bryk, 2002). This is because the estimates will confound within-subject and between-subjects associations.

In hypothetical Study 2, a within-subject association exists when subjects give higher ratings of liking to individual classes that they consider more prestigious. A between-subjects association exists when subjects who consider the eight classes, on average, to be rather prestigious also tend to like the eight classes more on average. The between-subjects association could be a real psychological effect—the more individuals think that the classes they took will look good on their transcripts the more they like classes in general—or it could be a scale usage effect, a mood effect, or a spurious relationship caused by a third individual difference variable (e.g., desire to be challenged). Regardless of the origin of the between-subjects association, it is theoretically distinct from the within-subject association.

The importance of the distinction between within- and between-unit associations should not be underestimated. Numerous articles have been written about the "ecological fallacy," the false assumption that a relationship between two variables at one level (e.g., within units) is necessarily the same at a different level (e.g., between units; Brewer & Venaik, 2014). Given that the parameter estimate for the predictor is an amalgam of the two relationships when the predictor is uncentered or centered around its grand mean, researchers unknowingly commit the ecological fallacy when interpreting coefficients from models that do not use cluster-mean centering.[9]

Table 7
The Data from Hypothetical Study 2 in Long Format

Row	subject.ID	like	prestige	gender	age	item.ID
1	1	4	3	2	18	1
2	1	7	2	2	18	2
3	1	4	4	2	18	3
4	1	6	5	2	18	4
5	1	4	6	2	18	5
6	1	1	1	2	18	6
7	1	2	8	2	18	7
8	1	6	9	2	18	8
9	2	5	6	1	21	9
10	2	7	9	1	21	10
11	2	4	3	1	21	11
12	2	4	7	1	21	12
13	2	1	1	1	21	13
14	2	3	5	1	21	14
15	2	2	2	1	21	15
16	2	9	8	1	21	16
. . .	. . .	. . .	. . .	. . .	. . .	. . .
400	50	7	6	2	20	400

Note. Only the data from the first two subjects and the last row of the data file are shown. The full data file contains 400 lines.

Table 8
The LMEM and the R Script for Hypothetical Study 2 With One Continuous Within-Subject Variable ("Prestige")

$$like = \beta_0 + \beta_1 prestigeC + u_0 + u_1 prestigeC + e \quad (5)$$

```
d$prestigeC <- d$prestige -
  ave(d$prestige,d$subject.ID)
model1_2 <- lmer(like ~ 1 + prestigeC +
  (1 + prestigeC | subject.ID), data = d)
summary(model1_2)
Anova(model1_2, type = 3, test = "F")
```

Note. The predictor variable is being centered around each subject's own mean in order to avoid the confounding of between- and within-subjects effects.

[9] Enders and Tofighi (2007) discuss one situation in which it makes more sense to center the continuous within-subject predictor around its grand mean: when the goal is to examine the effect of a between-subjects variable while statistically controlling for the effects of a continuous within-subject variable. Note that in this situation, however, the within-subject variable is not the focus of the researchers' hypothesis. Additional information on within- versus between-unit associations can be found in Appendix B.

Related Analyses

The LMEM framework can also be used to examine the interaction of a continuous within-subject predictor with one or more other predictors. Tables 9 and 10 contain examples testing the interaction of a continuous within-subject predictor with one other predictor. In both examples, the other predictor is a dichotomous variable that varies either between (see Table 9) or within (see Table 10) subjects. Similar models can be estimated when the other predictor is continuous. In these cases, the other predictor should be cluster-mean centered if it varies within subjects, and grand-mean centered if it varies between subjects.

Note that the continuous within-subject predictor can be time, so that the model examines subjects' change over time. LMEMs are thus ideally suited to conduct growth-curve analysis (Liu, Rovine, & Molenaar, 2012), although readers should be aware that other data-analytic strategies exist (Kline, 2015). LMEMs can easily handle unequal time intervals between measurement moments and time intervals that differ from one subject to the next.

Extension to Multilevel Models

In all of the examples discussed so far, the predictor varied within subjects. Readers should be aware that everything said above (and in the rest of the article) also applies to studies in which a predictor varies within a higher-order unit or "cluster." This is the case, for example, when subjects are nested in groups, families, or classrooms. The so-called "multilevel models" are thus a specific case of LMEMs. Researchers working in the multilevel tradition sometimes use different terminology (e.g., level-1 and level-2 models), but statistically speaking, there are no differences between multilevel models and LMEMs (Gelman & Hill, 2006).

Examples of Multilevel Models

Consider a study in which researchers form discussion groups of four individuals and assign two of the individuals a high status (high prestige) and two of them a low status (low prestige). Afterward, all group members evaluate the extent to which they enjoyed the group discussion. The data file in long format will have one row per subject. The data can be analyzed with the LMEM described in Equation 2 and the R-script in Table 3 (with one minor change: `subject.ID` is replaced by `group.ID`). Subjects and groups are both random variables. Given that there is only one observation per subject, the LMEM contains no by-subject random effects. It does contain two random effects for group: One by-group random intercept (because the four data points from the same group are dependent) and one by-group random slope (because the predictor "prestige" varies within groups).

Table 9
The LMEM and the R Script for Study With One Continuous Within-Subjects Predictor (Prestige) and One Dichotomous Between-Subjects Predictor (Gender)

$$\text{like} = \beta_0 + \beta_1\text{prestigeC} + \beta_2\text{genderC} + \beta_3\text{prestigeC} * \text{genderC} + u_0 + u_1\text{prestigeC} + e \quad (6)$$

```
d$prestigeC <- d$prestige -
  ave(d$prestige,d$subject.ID)
d$genderC <- d$gender - 1.5
model_2b <- lmer(like ~ 1 + prestigeC * genderC +
  (1 + prestigeC | subject.ID), data = d)
summary(model_2b)
Anova(model_2b, type = 3, test = "F")
```

Table 10
The LMEM and the R Script for Study With One Continuous Within-Subjects Predictor (Prestige) and One Dichotomous Within-Subjects Predictor (Science)

$$\text{like} = \beta_0 + \beta_1\text{prestigeC} + \beta_2\text{scienceC} + \beta_3\text{prestigeC} * \text{scienceC} + u_0 + u_1\text{prestigeC} + u_2\text{scienceC} + u_3\text{prestigeC} * \text{scienceC} + e \quad (7)$$

```
d$prestigeC <- d$prestige -
  ave(d$prestige,d$subject.ID)
d$scienceC <- d$science - 1.5
model_2c <- lmer(like ~ 1 + prestigeC * scienceC
  + (1 + prestigeC * scienceC|subject.ID), data = d)
summary(model_2c)
Anova(model_2c, type = 3, test = "F")
```

Consider another study in which students nested in classrooms rate the extent to which they like school-related activities. In addition, the researchers assign each student a prestige score based on the socioeconomic status (SES) of his or her parents. The data from this study can be analyzed with the LMEM and the R script shown in Table 8, with one minor change: all instances of `subject.ID` are replaced by `classroom.ID`. The LMEM contains two by-classroom random effects, the random intercept and the random slope for prestige.

Like before, the coefficient associated with the predictor (prestige) describes the within-classroom association between the predictor and the outcome variable only if the predictor is centered around each classroom's own mean. Here, the coefficient tells us whether higher SES students enjoy school-related activities more (or less). This within-classroom association is likely to differ from the between-classroom association, that is, the extent to which classrooms with a high percentage of high SES kids also tend to have a high percentage of kids who enjoy school-related activities. If the predictor is uncentered or centered around the grand mean, its coefficient is an uninterpretable amalgam of both types of associations (Raudenbush & Bryk, 2002). Researchers would commit the ecological fallacy if they attempted to interpret this coefficient. See Appendix B for additional information on this topic.

Should Higher-Order Units in Nested Designs be Treated as Fixed Variables?

In certain disciplines (e.g., economics), researchers often treat the higher-order unit as a fixed rather than a random variable. In the study mentioned above—researchers examine the association between prestige (SES) and liking for school activities among students who are nested in classrooms—this approach would consist of running a standard GLM in which students is the unit of analysis and in which the outcome variable is regressed on the predictor and M-1 contrast codes (M being the number of class-

Fn10 rooms included in the study).[10] In a recent article, McNeish and Kelley (2017) compare the two data-analytic approaches and draw some general conclusions.

First, both approaches are clearly better than ignoring the nonindependence caused by the higher-order unit, which leads to an increased type-I error rate.

Second, the "fixed effects model" approach treats the higher-order unit as a fixed variable. By using this approach, the researchers are thus assuming that data have been gathered from all the levels of the higher-order unit that are of interest. They also accept the premise that the results of their study do not necessarily generalize to other higher-order units that were not included in the study. Such an approach may be acceptable when the goal is to solve a company-specific problem by measuring employees nested in departments, when all departments of the company have been included in the study, and when the researchers do not want to generalize their findings to other departments (e.g., departments in other companies). It may also be acceptable if there are only a few higher-order units and the statistical power of the LMEM would be unreasonably low. Such an approach is questionable, however, when researchers form discussion groups in the lab and want to generalize their results to discussion groups in general.

Third, most multilevel experts seem prefer the LMEM approach, for the reasons outlined above (Gelman & Hill, 2006; AQ: 6, 7 personal communications from Bates, 2016; Snijders, 2015). The "fixed effect model" approach may be a defensible data-analytic strategy when the number of higher-order units is small (Snijders & Bosker, 2012, mention "less than 20") and when the number of lower-order units is large (i.e., when there are many observations per higher-order unit; see McNeish & Stapleton, 2016).

Special Case—Only One Observation per Cell

In all of the examples above involving a dichotomous predictor, there were multiple data points per higher-order unit and per level of the predictor. For example, in the study described in Table 1, each subjects judged two classes per prestige level, and in the first example in the previous section, each discussion group contained two members at each of the two prestige levels. What happens when the design is such that there is only one observation per cell? Such studies can easily be analyzed with the LMEM framework, but some (minor) adjustments are necessary. As a concrete example, let's consider a study in which each subject evaluates his or her liking for two classes, one high-prestige class and one low-prestige class. The data from this study can be entered in wide format and analyzed with a paired samples *t* test. Alternatively, they can be entered in long format and analyzed with a LMEM (both analyses will yield the same result). The equation for this LMEM is identical to the one in Equation 2: It contains two fixed effects and three elements in the "complex error term" (two random effects and random error).

It turns out, however, that two of the elements in the error term are confounded with each other in such a study: The by-subject random slope for prestige and the random error. Although the two sources of error exist in reality, they cannot be mathematically separated. If a prank-loving collaborator secretly introduced random error into your data file, you wouldn't know if the observed deviations from the model predictions are random error or caused by the fact that subjects vary in their relative preference of one type of class over the other. Likewise, if you did a replication of the first study but you inadvertently recruited subjects who vary more in their relative preference of one type of class over the other than the subjects in the original study, you wouldn't know if the larger error term in the replication study is due to a change in the random slope of prestige (between-subjects variation in relative preference) or due to a change in random error. In order to make the confound explicit we use the bracket notation introduced by Judd, Westfall, and Kenny (2017). The brackets indicate the variance components that are confounded with each other.

$$\text{like} = \beta_0 + \beta_1 \text{prestigeC} + u_0 + [u_1 \text{prestigeC} + e] \quad (8)$$

A similar confound exists in more complex designs. Consider a slightly modified version of a study presented earlier (see Table 6). Subjects now rate their liking for four classes: a high-prestige science classes, a high-prestige nonscience class, a low-prestige science classes, a low-prestige nonscience class. This study has a 2×2 fully within-subjects design and there is only one observation per subject and per cell. It can be shown that the by-subject random slopes are confounded with the random error. The question is, then, how to deal with this confound.

The best solution is to avoid this situation altogether by having more than one observation per cell. As data analysts, we want to be able to model both between-subjects variation (in how subjects are affected by the within-subjects predictor) and random error (i.e., measurement error and other random fluctuations). Once the random error is correctly quantified, the statistical model can remove it from the equation so we can get an accurate assessment of the extent to which the effect of the within-subjects predictor varies from one subject to the next. Multiple random errors cancel each other out, and multiple measurements lead to more reliable assessments of the construct under consideration. As a consequence, we should include multiple observations per cell of the within-subjects design whenever possible. This can easily be achieved by including more items, measuring each subject more than once per condition, or simply computing two scores from the outcome measure (e.g., even and odd items, first and second minute).

When a study does not allow for more than one observation per cell, researchers have three options to deal with the existing confound. The first option is to ignore the issue. Some statistics programs, like R, will estimate a LMEM and produce relevant output. Although the output for the random effects should not be interpreted—the software will make an attempt to estimate parameters for all variance components, even the ones that are confounded—the fixed part of the model can be trusted. The fixed parameters will provide unbiased estimates of the population effects and will replicate the results of the repeated measures ANOVA with data in wide format.[11] Fn11

The second option is to set the variance for the random error to a very small value, let's say .00001 (possible with blmer in R and

[10] Mathematically equivalent approaches are to regress the outcome variable (a) on the intercept, the predictor, and M-1 dummy codes; or (b) on the predictor and M dummy codes (and to remove the intercept).

[11] The data analysis program will most likely generate an error message with the output. This error message can be ignored. In R, it is possible to suppress the error message by adding `control = lmerControl(check.nobs.vs.nRE = "ignore")` to the model statement as the last element in the parenthesis.

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

the PARMS argument in SAS, e.g.), or to set it to zero. Although none of the major data analysis programs currently allow users to set the variance for the random error to zero, such a (highly desirable) option may become possible in the future. A third option is to transform the data file to wide format and to analyze the data with a repeated-measures ANOVA.

A fourth option is available if there is only one within-subject predictor in the model. In this case, the researchers can simply delete the by-subject slope for the predictor (Barr et al., 2013, p. 275). The model will correctly estimate the parameter for the fixed slope, and its inferential test will have a type-I error rate of 5%. Note that this option is not available when there are two or more within-subjects predictors. It has been suggested that in such a case it suffices to delete the random slope for the highest-order interaction term in the LMEM. Our simulations show that such a LMEM will *not* reproduce the results of the repeated measures ANOVA with data in wide format.[12] Fn12

Restricted Maximum Likelihood

It should be noted that LMEMs use an estimation procedure called "Restricted Maximum Likelihood" (ReML), and not, as the standard ANOVA and regression analysis, "Ordinary Least Squares" (OLS). ReML is an iterative process in which the parameter estimates are progressively modified to maximize the "log likelihood function." At each step, the computer program estimates the parameters and determines the likelihood of having obtained the data at hand if the population parameters really had those values. In the following step, it changes the parameter estimates based on certain algorithms and tests if the new values yield an even greater likelihood (Demidenko, 2013). The iterative process stops when the log likelihood function can no longer be maximized by further changes to the parameter estimates. As opposed to Maximum Likelihood, ReML produces unbiased estimates of variance and covariance parameters. ReML and OLS yield identical results only in so-called "simple LMEMs," in which all within-subject variables are categorical and subjects are the only source of nonindependence.

The use of ReML as the estimation procedure has a number of implications that the user of LMEMs should be aware of. Except in simple LMEMs, the final model will likely have denominator degrees of freedom with decimals. They should be reported as such, for example, $F(1, 54.17) = 4.86$, $p < .04$. There is no agreement among statisticians about the best way to compute the appropriate *df*s (Baayen, Davidson, & Bates, 2008). Following the lead of Judd et al. (2012) we suggest using the Kenward-Roger approximation to compute the *df*s (Kenward & Roger, 1997). This approach uses the equally acceptable Satterthwaite approximation (the default in SAS and the only method used in SPSS), but will rescale the F ratio and compute the degrees of freedom in a way that results in a better approximation to an appropriate F distribution. The Kenward-Roger method is available in most standard data analysis programs. It is the default in R. ReML and the Kenward-Roger approximation both require large sample sizes to yield stable estimates. It is generally advised to have at least 200–300 observations (>500 are considered ideal; Raudenbush & Bryk, 2002).

Throughout this article, we provide the R scripts allowing researchers to compute a F-statistic with Kenward-Rogers degrees of freedom. Some researchers use a likelihood ratio test statistic that yields a chi-square value (e.g., Bates, Kliegl, Vasishth, & Baayen, 2015a). Both methods are acceptable and yield comparable results. There are some studies suggesting that the Kenward-Roger statistic outperforms the likelihood ratio test (in terms of maintaining the nominal alpha level), especially in small samples (Kenward & Roger, 1997; Kuznetsova, Brockhoff, & Christensen, 2015). Only the results of the Kenward-Roger F-statistic will exactly reproduce those of the repeated measures ANOVA.

One advantage of ReML estimation is that it can easily handle missing values. Whereas the repeated measures ANOVA will delete subjects if they have one or more missing values, a LMEM can derive parameter estimates and compute inferential statistics even when the data are incomplete (Rasbash et al., 2000). It will simply use the data that are available and take into account that the fact that the relationship between the predictor and outcome has been estimated more reliably for certain subjects (with complete data) than for other subjects (with incomplete data). An implication is that LMEMs can appropriately analyze data from studies in which different subjects provided a different number of observations for each level of the predictor variable (so-called "unbalanced repeats"). Whereas repeated measures ANOVAs limit themselves to computing total scores by averaging across multiple data points in the same cell of the design, LMEMs also take into account the reliability of each of the total scores (which is determined by the number and the variability of the data points).

Not all experts agree about the minimum number of levels that a random variable should have before one can include random effects for it. Does it make sense to include a by-subject random slope in a study in which six subjects each evaluate 100 stimuli, or a by-school random slope in a study in data from 1,000 students nested in four schools are collected? Raudenbush and Bryk (2002) suggest that one should have at least 10 levels. Stegmueller (2013) argues that fewer levels are acceptable as long as one is interested only in the fixed effects and the model does not contain any interactions of variables of different type (e.g., the interaction between a within- and a between-subjects variable). According to our understanding it is impossible to suggest clear guidelines. In order to maintain the type-I error rate at 5%, it is necessary to include by-unit random effects whenever the unit causes nonindependence in the data, regardless of the number of levels. It has been shown that the type-I error rate inflation is higher when the number of levels is small (Judd et al., 2012) and persists even if the random effect for the predictor under consideration has a near-zero variance (Barr et al., 2013). The problems of a small number of levels are low statistical power (type-II errors) and the instability of the observed effects. These are serious problems that can be addressed in a variety of ways—for example, include a larger number of levels (even if this implies a smaller total number of observations), sacrifice generalizability by treating a random variable as if it were fixed (see section on Multilevel Models, Appendix C, and McNeish & Stapleton, 2016)—but simply ignoring a random variable that causes nonindependence is not a viable option.[13] Fn13

[12] The R script for the simulations is available upon request.

[13] See the section Deciding on the LMEM to be Estimated for more information on the inclusion of random effects with a variance component that is not reliably different from zero.

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Multiple Sources of Nonindependence

In many psychological experiments, subjects are exposed to the same set of items, that is, they may view the same words, sentences, pictures, or avatars, or they may rate the same products, faces, art works, or individuals. Alternatively, subjects clustered in groups, families, classrooms, or counties may each provide multiple data points. Such studies contain two sources of nonindependence: Some responses may be more similar because they were made by the same subject, other responses may be more similar because they concern the same item (i.e., different subjects rating the same item), and yet other responses may be more similar because they were made by subjects in the same group.

Statistical analyses of such data have to take this double source of nonindependence into account. Judd et al. (2012) have shown that failure to do so leads to an increased type-I error rate, sometimes as high as 60%. By averaging across items belonging to the same category, we are not taking into account the nonindependence due to items. By averaging across subjects belonging to the same group, we are ignoring the variability between subjects. No data analysis technique belonging to the general linear model (e.g., ANOVA, regression analysis) can effectively deal with multiple sources of nonindependence. Only LMEMs with the appropriate random effects can achieve this goal (Baayen et al., 2008).

Readers should be aware that "items" and "groups" are not the only possible second source of nonindependence (subjects is usually the first source). When subjects interact with one of several confederates—where half are European American and half are African American—the design calls for the inclusion of by-confederate random effects. If a researcher runs a study at multiple, randomly selected locations across the country and wants to generalize her findings to the entire U.S., then she should include by-location random effects. In cross-cultural research with respondents from many countries, it not only makes sense to include the appropriate by-country random effects but also to distinguish the within-country associations from the between-country associations. To summarize, researchers should carefully examine whether their studies contain possible sources of nonindependence and account for the dependence in the data analyses by specifying the appropriate LMEM. Failure to do so will lead to increased type-I error rates.

When there are multiple sources of nonindependence, it can become difficult to decide which random effects should be included in the analysis. We will address this issue in the next sections of this article. Readers should be aware that the inclusion of random effects is an active area of research, and that new articles on this issue are published on a regular basis (e.g., Baayen, Vasishth, Bates, & Kliegl, 2017; Winter & Wieling, 2016). Even worse, there is some disagreement between experts regarding the choice of the appropriate random effects structure.

Fortunately, however, all experts agree about the initial steps of the data analysis. The first step consists of determining the so-called "maximal random effects structure" that the design of the study calls for. As we will see, this maximal random effects structure includes all random effects that we might want to include based on the characteristics of the design, that is, based on whether the predictors vary within or between levels of the random variables that cause nonindependence in the data. The second step consists of estimating a LMEM with this maximal random effects structure. As we will see, however, this is not always possible because some LMEMs are so complex that the iterative estimation procedure fails to converge. In such a case it is necessary to progressively simplify the random effects structure until convergence can be achieved.

In the next sections, we will discuss each of these first two steps in detail. Our discussion, we will focus on two random variables that cause nonindependence in the data. The first random variable is always subjects. We will refer to the second random variable as "items," but readers should know that this term may refer to any random variable that is a second source of nonindependence (e.g., stimuli, targets, confederates, locations, groups, classrooms). In the following section, entitled Deciding on the LMEM to be Estimated, we will come back to the disagreement among experts regarding the analyses that should be performed after the second step. We will see that the proposed analyses do not fundamentally differ from each other.

How to Determine the Maximal Random Effects Structure

Random Intercepts

Nonindependence is accounted for by including the appropriate random intercepts in the maximal random effects structure. The first rule in Table 11 identifies when random intercepts are required. When the "unit" under consideration is subjects, then this rule says: If there is nonindependence due to subjects, then the maximal random effects structure should include a by-subject random intercept. In other words, whenever a given subject provides multiple data points, a by-subject random intercept should be included. When the term "unit" refers to items, then the rule translates into: If there is nonindependence due to items, the maximal random effects structure should include a by-item random intercept. Said differently, when all (or some of the) subjects evaluate the same set of items, a by-item random intercept should be specified. Note that both conditions can be satisfied at the same time. When the same set of subjects provides responses to the same set of items, then the maximal random effects structure should include both a by-subject random intercept and a by-item random intercept. T11

Note that the first rule is merely an extension of a topic discussed earlier. Both subjects and items have been randomly se-

Table 11
A Set of Simple Rules Regarding the Types of By-Subject and By-Item Random Effects That Should be Included in the Maximal Random Effects Structure (Adapted from Barr et al., 2013)

First rule: If a unit causes nonindependence then a by-unit random intercept is required.
Second rule: In general, a within-unit predictor requires a by-unit random slope, whereas a between-unit predictor does not.
Third rule: It is advised to include a by-unit random slope for interactions when all factors comprising the interaction are within-units.

Table 12
The LMEM and the R Script for a Study With One Dichotomous Predictor (Prestige) That Varies Between-Subjects and Between-Items

$$\text{like} = \beta_0 + \beta_1 \text{prestigeC} + u_0 + v_0 + e \quad (10)$$

```
d$prestigeC <- d$prestige - 1.5
model_3a <- lmer(like ~ 1 + prestigeC +
  (1|subject.ID) + (1 | item.ID), data = d)
summary(model_3a)
Anova(model_3a, type = 3, test = "F")
```

Note. All the subjects in the same experimental condition rate the same set of items. The LMEM contains two fixed effects (the fixed intercept β_0 and the fixed slope β_1), two random effects (the by-subject random intercept u_0 and the by-item random intercept v_0), and the random error (e)

lected from a larger pool of possible exemplars, so they are both random variables. In all the examples discussed in earlier sections of this article, there were multiple data points per subject but only one data point per item. However, when there are multiple data points for each subject and multiple data points for each item, these two sources of nonindependence have to be taken into account by adding both a by-subject random intercept and a by-item random intercept.

Imagine a study with 100 subjects, half of whom rate their liking for a set of 30 low-prestige cars, whereas the other half evaluate a set of 30 high-prestige cars. In this study, the variable `subject.ID` has 100 levels, whereas the variable `item.ID` has 60 levels. The data contain nonindependence due to subjects because each subject provides multiple ratings, thus requiring a by-subject random intercept. But the data also contain nonindependence due to items because each item is being evaluated by multiple subjects, creating the need for a by-item random intercept. Table 12 provides the equation for the LMEM and the R script when all random effects are included in the analysis. Note that the maximal random effects structure for the LMEM does not include a by-subject random slope for prestige because prestige varies between-subjects. We will provide more details about this point in the following section.

There is one exception to this rule. When the unit under consideration is fully confounded with the predictor, then no by-unit random intercept is required. Imagine a study in which half of the subjects evaluate one low-prestige car and the other half of the subjects evaluate one high-prestige car. Each subject provides only one rating, so subjects do not create any nonindependence and no by-subject random effects are included in the LMEM. Although there is nonindependence due to items—each item is evaluated by multiple subjects—we still would not include a by-item random intercept, because the random variable `car.ID` is fully confounded with the fixed manipulated predictor `prestigeC`. Given that the resulting LMEM contains no random effects, it is equivalent to an independent samples *t* test.

Random Slopes

How do we know which random slopes to include in our maximal random effects structure? Barr et al. (2013) suggested the second rule shown in Table 11. We want to include a by-subject random slope for any predictor that varies within subjects (but not when it varies between subjects). Likewise, we will want to include a by-item random slope for any predictor that varies within items (but not when it varies between items).

In order to illustrate this rule, let's imagine a study in which subjects rate their liking for eight low-prestige cars and eight high-prestige cars. All subjects rate the same set of 16 cars. Before Judd et al. (2012), some of us may have been tempted to enter the data in wide format, compute two scores per subject (the average rating for the eight low-prestige cars and the average rating for the eight high-prestige cars), and analyze these scores with a repeated-measures ANOVA or a paired-samples *t* test. We now know that such an analysis would produce biased standard errors and thus an increased type-I error rate. Given that the data are nonindependent due to multiple ratings per subject and multiple ratings per item, a LMEM is the appropriate data-analytic strategy. The maximal random effects structure of the LMEM contains three random effects (see Table 13). Note that no by-item random slope for prestige is needed because prestige does not vary within items (each target car is either low or high in prestige).

The second rule can also be applied to studies in which the predictor varies between-subjects but within items. Imagine a study in which subjects evaluate the same set of 20 cars (of medium prestige) but they do so in one of two between-subjects conditions. Half of the subjects rate their liking for the 20 cars after having seen five highly prestigious cars (high prestige context), whereas the other half of the subjects were previously exposed to five cars of very low prestige (low prestige context). Again, this study contains two sources of nonindependence, subjects and items, requiring two random intercepts. The predictor "prestige" varies between-subjects (each subject is in only one prestige context condition), but varies within-items (each of the 20 target cars is in the high prestige context for certain subjects but in the low-prestige context for other subjects). The maximal random effects structure for the appropriate LMEM contains three random effects (see Table 14). Given that prestige (context) varies between-subjects, there is no need to include a by-subject random slope for prestige.

The final case covered by the second rule is where the predictor varies both within-subjects and within-items. Imagine a study in which subjects are shown the same set of 20 hybrid cars. For every subject, 10 cars are shown in a high-prestige context and 10 cars

Table 13
The LMEM and the R Script for a Study with One Dichotomous Predictor (Prestige) That Varies Within Subjects but Between Items

$$\text{like} = \beta_0 + \beta_1 \text{prestigeC} + u_0 + u_1 \text{prestigeC} + v_0 + e \quad (11)$$

```
d$prestigeC <- d$prestige - 1.5
model_3b <- lmer(like ~ 1 + prestigeC +
  (1 + prestigeC|subject.ID) + (1|item.ID),
  data = d)
summary(model_3b)
Anova(model_3b, type = 3, test = "F")
```

Note. All subjects rate the same set of items. The LMEM contains two fixed effects (the fixed intercept β_0 and the fixed slope β_1), three random effects (the by-subject random intercept u_0, the by-subject random slope for prestige, and the by-item random intercept v_0), and the random error (e).

are shown in a low-prestige context. In addition, the context in which a given car is shown is counterbalanced across subjects. The predictor "prestige" now varies within-subjects (each subjects sees some cars in a high prestige context and other cars in a low-prestige context) and within-items (the same car will be shown to certain subjects in a high prestige context and to other subjects in a low-prestige context). The design therefore calls for four random effects: a by-subject random intercept, a by-subject random slope for prestige, and a by-item random intercept, and a by-item random slope for prestige.

In the examples discussed in the previous paragraphs, the predictor (prestige) was always dichotomous (high vs. low). The same rules from Table 11 apply to continuous predictors that vary either between or within subjects and either between or within items.

Sometimes items are nested in subjects or individuals are nested in groups. As before, the same rules apply. Imagine students nested in universities who evaluate two high-prestige classes and two low-prestige classes. The predictor, prestige, varies within-subjects and within-universities, and the maximal random effects structure would thus have four random effects: a by-subject random intercept, a by-subject random slope for prestige, and a by-university random intercept, and a by-university random slope for prestige.

There are two caveats to the second rule shown in Table 12. First, a by-item random effect is included in the maximal random effects structure only if there are multiple observations per level of the predictor. For example, if all subjects rate the same set of two cars, one low-prestige car and one high-prestige car, then it does not make sense to include any by-item random effects, because variability between items is fully confounded with variability between conditions. Second, by-item random effects are included only if subjects judge the same set of items. If different subjects provide ratings about different items (as in Studies 1 and 2 mentioned in the initial sections of this article), there is no nonindependence due to items that has to be taken into account.

Random Effects for Interactions

With regard to interactions, Barr et al. (2013) suggest the third rule provided in Table 11. When all predictors of an interaction vary within-subjects, a by-subject random slope for the interaction term should be included in the maximal random effects structure.

Table 14
The LMEM and the R Script for a Study With One Dichotomous Predictor (Prestige) That Varies Between-Subjects but Within-Items

$$\text{like} = \beta_0 + \beta_1\text{prestigeC} + u_0 + v_0 + v_1\text{prestigeC} + e \quad (12)$$

```
d$prestigeC <- d$prestige - 1.5
model_3c <- lmer(like ~ 1 + prestigeC +
  (1|subject.ID) + (1 + prestigeC|item.ID),
  data = d)
summary(model_3c)
Anova(model_3c, type = 3, test = "F")
```

Note. All subjects rate the same set of items. The LMEM contains two fixed effects (the fixed intercept β_0 and the fixed slope β_1), three random effects (the by-subject random intercept u_0, the by-subject random slope for prestige, and the by-item random intercept v_0), and the random error (e).

Table 15
The LMEM and the R Script for a Study With One Dichotomous Predictor ("Party Affiliation"), One Continuous Predictor ("Openness") and Their Interaction

$$\text{eval} = \beta_0 + \beta_1\text{affil} + \beta_2\text{open} + \beta_3\text{affil} * \text{open} + u_0 + u_1\text{affil} + v_0 + v_1\text{open} + e \quad (13)$$

```
d$affilC <- d$affiliation - 1.5
d$openC <- d$openness - mean(d$openness)
model_11 <- lmer(like ~ 1 + affilC * openC +
  (1 + affilC|subject.ID) + (1 + openC|item.ID),
  data = d)
summary(model_11)
Anova(model_11, type = 3, test = "F")
```

Note. "Party affiliation" varies within-subjects and between-items, whereas "openness" varies between-subjects and within-items. The parameters β_0, β_1, β_2, and β_3 estimate the fixed effects, the parameters u_0 and u_1 estimate the by-subject random effects, and the parameters v_0 and v_1 estimate the by-item random effects.

Likewise, when all predictors of an interaction vary within-items, researchers should include a by-item random slope for the interaction term. This rule, in addition to the previous two, allows us to determine the maximal random-effects structure for more complex designs.

Consider a study in which subjects evaluate 20 well-known politicians (10 Democrats and 10 Republicans, "party affiliation," a dichotomous predictor). The researchers also measure subjects' level of openness to experience ("openness," a continuous predictor). The fixed effects structure is relatively easy to determine: it contains the overall intercept, the effect for party affiliation, the effect for openness, and the interaction between the two predictors (four effects in total). To find the maximal random-effects structure, one can use the rules in Table 11. There is nonindependence due to subjects and items, and we thus want to include a by-subject random intercept and a by-item random intercept (first rule). Party affiliation varies within-subjects but between-items. We thus need to specify a by-subject random slope for party affiliation (second rule). Openness to experience varies between-subjects but within-items. The design thus calls for a by-item random slope for openness (second rule). It is not the case that both predictors vary within-subjects, and it is also not the case that they both vary within-items. We thus do not have to include any random effects for the interaction term (third rule). The maximal random effects structure contains four random effects: a by-subject random intercept, a by-subject random slope for party affiliation, a by-item random intercept, and a by-item random slope for openness to experience. The full model and the R script for the LMEM with the maximal random effects structure is shown in Table 15.[14] T15, Fn14

As an additional example, consider a study in which male and female subjects complete a lexical-decision task. The 80 target words are either agentic or communal ("word type") and they are either positive or negative ("word valence"). The researchers predict a three-way interaction. The fixed effects structure contains

[14] If each subject evaluated only two politicians, one Democrat and one Republican, then all by-item effects would drop out of the LMEM, because `item.ID` would be fully confounded with the predictor `affil.ID`.

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

T16 eight effects (see Table 16). The maximal random effects structure requires two random intercepts because both subjects and items cause nonindependence. It further requires four random slopes: one by-subject random slope for word type (because word type varies within-subjects and between-items), one by-subject random effect for word valence (because word valence varies within-subjects and between-items), one by-subject random slope for the word type by word valence interaction (because both predictors vary within-subjects), and one by-item random slope for subject gender (because subject gender varies between-subjects and within-items).

The same reasoning about random effects for interaction terms can be applied to continuous predictors and studies in which one random variable is nested in another random variable (e.g., items nested in subjects or students nested in classrooms).

Note that the rules in Table 11 do not apply to designs in which subjects judge each of the items multiple times in different conditions, that is, designs in which both subjects and items are crossed with the predictor(s). A typical example for such a design would be a study in which subjects are exposed to the same set of target words twice, once in lower case letters and once in upper case letters, and the researchers want to examine if subjects react to the words differently (e.g., faster, more positively) if they are presented in one case rather than the other. Readers are advised to consult more specialized publications to determine the maximal random effects structure for these types of designs (e.g., Judd et al., 2017).

How to Address Convergence Problems

After the maximal random effects structure for the study has been determined, the next step is to estimate the corresponding LMEM. Unfortunately, the iterative ReML procedure to derive the parameter estimates does not always "converge" (i.e., the numerical optimization algorithm cannot reliably determine the maximum of the log-likelihood function). Such problems typically occur when a lot of parameters are being estimated in the LMEM or when there are only a few data points in one or more cells of the design.

Table 16
The LMEM and the R Script for a Hypothetical Study With Three Dichotomous Predictors Gender, Word Type, and Word Valence

$$\begin{aligned} \text{rt} = {} & \beta_0 + \beta_1 \text{w.type} + \beta_2 \text{w.valence} + \beta_3 \text{gender} + \beta_4 \text{w.type} * \text{w.valence} \\ & + \beta_5 \text{w.type} * \text{gender} + \beta_6 \text{w.valence} * \text{gender} \\ & + \beta_7 \text{w.type} * \text{w.valence} * \text{gender} + u_0 + u_1 \text{w.type} + u_2 \text{w.valence} \\ & + u_3 \text{w.type} * \text{w.valence} + v_0 + v_1 \text{gender} + e \end{aligned} \quad (14)$$

```
d$genderC <- d$gender - 1.5
d$w.typeC <- d$w.type - 1.5
d$w.valenceC <- d$w.valence - 1.5
model_12 <- lmer(rt ~ 1 + genderC * w.typeC *
  w.valenceC + (1 + w.typeC * w.valenceC|
  subject.ID) + (1 + genderC|item.ID), data = d)
summary(model_12)
Anova(model_12, type = 3, test = "F")
```

Note. "Word type" and "word valence" vary within-subjects and between-items, whereas "gender" varies between-subjects and within-items. The parameters β_0 to β_7 estimate the fixed effects, the parameters u_0 to u_3 estimate the by-subject random effects, and the parameters v_0 and v_1 estimate the by-item random effects.

Although not absolutely necessary, it is helpful to understand the number of parameters that are being estimated in a given LMEM, to anticipate potential problems in the estimation procedure, often referred to as "convergence problems" (Bates et al., 2015a). A typical LMEM estimates one parameter for each fixed AQ: 8 effect and one parameter for each random effect. It also estimates one parameter for the variance of the residuals (the *e*'s). Finally, the LMEM estimates one parameter for every possible covariance between all random effects belonging to the same unit. See Appendix D for some concrete examples.

Regardless of the number of parameters being estimated, researchers should first run a model with the maximal random effects structure. If this model converges, they can then move on to the third step, described below. If this model fails to converge, however, they can use the "remedies" listed in Table 17 (most of T17 which are based on Barr et al., 2013, and Winter, 2016). The remedies are ordered hierarchically, so that researchers should try one remedy at a time, reestimate the model, and move down the list only if this remedy does not solve the convergence problem. Many of the suggested remedies are self-explanatory. We will comment on a subset of them in the following paragraphs.

The first remedy suggests checking for outliers and violations of model assumptions (e.g., linearity). Be aware, however, that these types of checking procedures have not yet been widely implemented in software packages (but see Loy & Hoffmann, 2014). AQ: 10

It is generally advised to center all predictors regardless of whether they participate in an interaction term or not, as centering reduces multicollinearity in the random effects structure (Remedy 5). Maximum likelihood estimation has difficulties with highly correlated parameters. As mentioned above, it makes sense in most cases to code dichotomous within-subject predictors as −.5 and +.5 and to center continuous within-subject predictors around each subject's own mean.

Most data analysis programs allow researchers to increase the number of iterations, to change the numerical optimization procedure, and to provide better starting values (Remedies 6, 7, and 8). Bates et al. (2015a) report a LMEM that converged after 39,004 iterations and discuss different numerical optimization procedures. Subject matter knowledge, results from past studies, and reports in the literature are good sources for determining appropriate starting values. Sometimes it is possible to estimate a simplified model (e.g., without covariances among random effects) and to use the parameter estimates of this model as starting values for the full LMEM that contains the maximal random effects structure. An in-depth discussion of these three remedies in iterative parameter estimation goes beyond the scope of this article (but see Bates et al., 2015b).[15] fn15

Rescaling the predictor variables, the outcome variable, or both may address a convergence problem (Remedy 9). Certain algo-

[15] In R, the number of iterations or the optimization procedure can be changed with the function `lmerControl`, starting values can be specified using the function `start`.

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Table 17
A List of 20 Remedies That Can be Used to Achieve Convergence of LMEMs

Preventive measures to avoid convergence failures:
1. Include as many subjects and as many items as possible in your study.
2. Include heterogeneous subjects and heterogeneous items.

Nonintrusive remedies to address convergence failures:
3. Check your data (case analysis, distributional assumptions). If necessary apply a transformation to the predictor(s) or the outcome variables.
4. Check for model misspecifications (e.g., does the model contain a by-unit random slope for a predictor that varies between units?).
5. Make sure all predictors are centered (reduces multicollinearity).
6. Increase the number of iterations.
7. Change the numerical optimization procedure that is used to maximize the log likelihood function.
8. Give the model better starting values.
9. Rescale the predictor(s) or the outcome variable.
10. Check whether the nonconvergence is due to the presence of a few subjects (or items) with a small number of observations in particular cells. If yes, consider imputing data or removing the problematic subjects (or items).
11. Remove random effects for covariates (as long as the interactions between the covariates and the factors of interest are not in the model).
12. Check whether your model can be simplified: If the hypothesis is a X1 by X2 interaction, then it may not be necessary to include X3 and X4 in the model (at least initially). Or remove all but one predictor in a set of highly correlated predictors.

Classic remedies to address convergence failures:
13. If the goal is to test two fixed effects, X1 and X2, but not their interaction, estimate two LMEMs, one with both fixed effects but only the random slope for X1 (to test X1), and one with both fixed effects but only the random slopes for X2 (to test X2).
14. If you have a design with two or more within-unit predictors and your hypothesis concerns the interaction, remove the by-unit random slopes for the within-unit predictors and the lower-order interactions, but do *not* remove the by-unit random slope for the highest-order interaction(s) between the within-unit predictors (Barr, 2013, *Frontiers*).
15. Selectively remove covariances among random effects: Start out by removing covariances of predictors that are not directly related to your hypotheses (X3 and X4 mentioned in remedy #11 if you have decided to keep these predictors in the model). Continue to remove covariances that you suspect to be close to zero anyway. Finally, remove all covariances among random effects.
16. Remove some or all of the random intercepts for which there are also random slopes in the model. Do not remove the random slopes. Warn your readers that you have estimated a model without random intercepts.
17. Perform two separate LMEMs—one with subject as the unit of analysis (with maximal by-subject random effect structure, but no by-item random effects) and one with item as the unit of analysis (with maximal by-item random effect structure, but no by-subject random effects) – and apply the $F_1 \times F_2$ logic (both have to be significant at $p < .05$).

Corrective remedies with major shortcomings:
18. Run the analyses and compute Clark's (1973) min-*F'* statistic. Be aware that this test is seriously underpowered. If min-*F'* is significant you can be confident that your result is not due to an inflated type-I error rate.
19. Run a LMEM or a repeated-measures ANOVA in which subjects is the only random variable (and thus ignore the nonindependence due to items). Warn your readers that your type-I error rate may be as high as 60% (see Table 2 in Judd et al., 2012).
20. Estimate a LMEM in which you keep the random intercepts but not the random-slopes. Warn your readers that your type-I error rate may be as high as 80% (see Table 5 in Barr et al., 2013).

Note. Each remedy is likely to affect parameter estimates to a greater extent than the previous one. It is thus advised to use a certain remedy only if all previous remedies have proven to be ineffective.

rithms do not perform well when the covariance parameters are on a different scale. Rescaling the effects may help (e.g., recode one of the dichotomous within-subject predictors into −10 and + 10). Extremely large or extremely small data values on any of the variables can cause convergence problems because of internal tolerances built in the data analysis software (Kiernan, Tao, & Gibbs, 2012).

LMEMs struggle with cells in the design that contain no or few data points. They also have a hard time converging if certain subjects or items have a lot of missing data (e.g., a single data point in a cell for which the LMEM expects multiple data points). Solutions for these problems are imputing missing data, removing problematic subjects/items, or simplifying the design that is being analyzed (Remedy 10).

In his 2013 *Frontiers* article, Dale Barr (2013) showed that as long as the random effect for the higher-order interaction is in the model, the random effects for the main effects and the lower-order interactions can be removed without an increase in type-I error rate if one's hypothesis is about the higher-order interaction term. This insight leads us to Remedy 14. Imagine a study with a $2 \times 2 \times 2 \times 2$ fully within-subjects design requiring us to estimate, among other parameters, 16 by-subject random effects and 120 covariances among by-subject random effects. Removing the main effects and the lower-order interactions from the by-subject random effects structure reduces the number of parameters to be estimated to three: the by-subject random intercept, the by-subject random slope for the four-way interaction, and the covariance among the two. A difference of 133 parameters! Note that the fixed-effects structure remains untouched: All main effects and interactions remain in the model.

Maximum likelihood estimation has problems with variance and covariance parameters that are zero or that are very close to zero. If one has theoretical or empirical reasons to believe that a certain covariance between two random effects should be close to zero, then it is acceptable to fix it to zero (Remedy 15). The *Expected Mean Square Calculator*, developed by Jake Westfall (jakewestfall.org), can help researchers decide which random effects are likely to be zero. See also Bates et al. (2015a) for a discussion of the parameters that are likely to be zero.[16] Fn16

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Remedy 16 may surprise certain readers. Simulations have shown that removing the random intercepts (falsely assuming that all subjects have the same average score on the outcome) leads to less biased parameter estimates than removing the random slopes for the predictor that is the focus of the researchers' hypothesis (falsely assuming that the effect of the within-subject predictor on the outcome is the same for all subjects; see Barr et al., 2013).[17] If one has multiple random intercepts to choose from, it usually makes more sense to remove the intercept with the smallest variance first. One can also compute the intraclass correlation (ICC) for subjects and for items, and then remove the random intercept of the unit that has the smallest ICC. This remedy assumes that the researchers' hypothesis concern a predictor (which is usually the case). In the rare case of a hypothesis about a fixed intercept it is advised to keep the random intercept. The general rule is: Researchers should never remove the random effect for a parameter for which they have a hypothesis, because the inferential test for a fixed parameter will maintain a 5% type-I error rate only if its corresponding random effect is in the model.

As an alternative to Remedies 15 (removing covariances among random effects) and 16 (removing random intercepts), which are both top-down approaches, Barr et al. (2013) also discusses two bottom-up approaches. One approach is to use "backward selection" and/or test slopes using the "best path" algorithm. The other approach is to use a "forward selection" with one of two selection criteria: (a) test the slopes in an arbitrary sequence or (b) at each step, test for the potential inclusion of *all* random effects not currently in the model, and include any that pass at a relatively liberal α-level (e.g., .20). Be aware that these two approaches are data-driven and that the result may not replicate with a different data set (see Barr et al., 2013, and Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017, for more details).

Note that estimating a model that ignores the nonindependence caused by a random variable (Remedy 19) and estimating a model with random intercepts but without random slopes (Remedy 20) are last on the list. Further note that several earlier remedies involve using data analysis strategies that belong to the family of general linear models, which by definition do not have convergence problems (e.g., Remedies 17 and 18). As a consequence, it is advised that researchers use these last two remedies only in exceptional circumstances. Until recently, models with a random intercept but without the random slopes were considered the gold standard in LMEMs (especially in multilevel modeling). We now know that these models have type-I error rates well above the acceptable level of 5%.

Deciding on the LMEM to be Estimated

In the previous sections, we have described the first two steps for the analysis of LMEMs. The first step was to determine the maximal random effects structure called for by the design of the study. The second step was to get the LMEM to converge while keeping the random effects structure as maximal as possible. What is the next step? Unfortunately the literature fails to provide a clear answer to this question. Experts like Barr et al. (2013) suggest the "keep-it-maximal approach" which consists of interpreting the first model that achieves convergence. Others, like Bates et al. (2015a) and Matuschek et al. (2017), propose a "model selection approach" in which the first model that converges is further simplified. The goal of the simplification process is to remove variance components (random intercepts, random slopes, or covariances among random effects) that have a near-zero variance.

Note that the "keep-it-maximal-approach" implies that there is no need to test whether a certain random effect (variance component) is statistically different from zero. Inferential tests of variance components are usually underpowered (especially when the random variable only has a small number of levels) and require specialized mixture chi-square tests (Savalei & Kolenikov, 2008).

The model selection approach consists of several highly similar strategies. Bates et al. (2015a) suggest conducting a principal component analysis of the random effects structure to identify the variance components that can be removed without a loss in goodness of fit of the model. The authors developed a function in R to perform this analysis. Matuschek et al. (2017) use a "backward selection heuristic" paired with a likelihood ratio test. The heuristic tests whether the model contains one or more variance components that are not reliably different from zero (using a liberal significance level of $\alpha = .20$). If yes, it removes the variance component with the smallest variance. The heuristic continues this process until the LMEM contains only variance components that differ reliably from zero.

There are several good arguments in favor of the model selection approach. Variance components with near-zero variance do not contribute to the goodness of fit of the model. If anything, they lead to "overparameterized models" that are not supported by the data, even if these models converge (Bates et al., 2015a). All other things being equal, including random slopes usually decreases the degrees of freedom and thus increases the standard errors of the fixed parameter estimates. As a consequence, the inclusion of random slopes that have a near-zero variance in the data may lead to an unnecessary decrease of statistical power (Matuschek et al., 2017).

There are also several arguments in favor of the keep-it-maximal approach. Barr et al.'s (2013) simulations have shown that the inclusion of random slopes for "critical predictors" (i.e., predictors related to the researchers' hypotheses) is necessary if one wants to keep the type-I error rate at 5%, even if these random slopes have a relatively small (and maybe nonsignificant) variance. Techniques like the above-mentioned backward selection heuristic are data driven and their outcome is influenced by randomness in the data. Even an exact replication of the same study may have a different random effects structure. Finally, it is unclear why over-

[16] In R, a covariance parameter between two random effects is set to zero by writing the two random effects in separate parentheses. Thus, the model `lmer(like ~ 1 + prestigeC + scienceC + (1|subject.ID) + (0 + prestigeC + scienceC|subject.ID)` estimates the by-subject random intercept, the by-subject random slope for prestige, the by-subject random slope for science, and the covariance between the two by-subject random slopes, but it does not estimate the covariances between the random intercept and each of the random slopes. It is possible to set all covariances to zero by adding an second vertical trait between the last random effect and the random variable that causes nonindependence, for example, `(1 + prestige||subject.ID)` instead of `(1 + prestige|subject.ID)`.

[17] In R, the variance parameter of a random intercept can be set to zero by replacing the 1 by a 0. Thus, in the model `lmer(like ~ 1 + prestigeC + (0 + prestigeC|subject.ID) + (0 + prestigeC|item.ID)`, both random intercepts are set to zero. Remember that omitting the "1" will not set the random intercept to zero, since `(prestigeC|subject.ID)` is equivalent to `(1 + prestigeC|subject.ID)`.

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

parameterized models, as long as they converge, are problematic. We are not aware of any simulations showing that such models lead to biased parameter estimates. The gain in statistical power when adopting the model selection approach seems to be relatively minor and limited to underpowered studies (Barr et al., 2013; Matuschek et al., 2017).

Although the keep-it-maximal approach and the model selection approach seem to be fundamentally opposed, their differences are in fact relatively minor. Most experts agree that the final LMEM needs to contain the random slope(s) for the predictor that is the focus of the researchers' hypothesis, regardless of the variance of this (these) random slope(s). They also agree that the presence of variance components with a near-zero variance in the random effects structure does not affect the goodness-of-fit of the model. Whether they are included or excluded will thus have a minimal effect on the significance tests of the fixed effects.[18]

Fn18

Statistical Power Versus Generalizability

Westfall, Kenny, and Judd (2014) demonstrated that the inferential test of the predictor (i.e., the fixed effect) is underpowered in most LMEMs. They coined the term "maximally attainable power" to describe the highest level of power that one can achieve with a given number of items. For example, in a study in which subjects evaluate both eight low-prestige and eight high-prestige cars (see Table 13), the statistical power for the condition effect will never exceed .40, even if the researchers collect data from thousands or even millions of subjects (assuming a medium effect size of $d = .5$ and an intermediate level of variability between subjects and between items). As a rule of thumb, Westfall et al. (2014) suggest to never use less than 16 items and 16 subjects, preferably more.

Researchers may be tempted to include only one item per treatment level. Paradoxically, reducing the number of items per treatment levels from, let's say, eight to one, drastically increases statistical power. This is because no by-item random effects are needed with one item per treatment level. As tempting as it may seem, this solution is unsatisfactory because of its negative effects on generalizability. When multiple items per treatment level are used, the study's conclusion can be generalized to the entire population of items from which they were drawn (assuming they were drawn randomly). When the researchers include only one item per level of the predictor, then the results hold only for the items included in the study.

To illustrate this point, consider two studies, one in which subjects rate one low- and high-prestige car (see section Special Case—Only One Observation per Cell) and another in which they rate eight low- and eight high-prestige cars (see Table 13). Although the first study is more powerful (because the random effects structure contains no by-item random effects), its conclusions cannot be generalized to other low- and high-prestige cars. After all, it could be that the results are limited to the two particular cars chosen by the researchers. A similar reasoning applies to experiments in which researchers have subjects interact with either a White confederate or a Black confederate and then measure behavioral outcome measures. If the researchers employ only one White and one Black confederate in the study, their results cannot be generalized to people's reactions to Whites and Blacks in general. If, however, the researchers randomly chose a group of White and Black individuals on campus to serve as confederates, then the results can be generalized to people's reactions to White and Black students in general. In the latter case, it is necessary to include appropriate by-confederate random effects. Confederate race varies between confederates. And if there is a between-subjects condition and the same confederate interacts with subjects in both conditions, then condition varies within confederates.

To summarize, deciding on one's experimental design is partly a trade-off between statistical power and generalizability. The ideal is to include more than 16 levels (preferably more) per random variable. When this is not possible due to constraints (e.g., not enough items exist, excessive length of the study, few schools gave permission to collect data), researchers should choose the design that has the greatest statistical power. Westfall et al. (2014) provide helpful advice on what this design is given the particular constraints faced by the researchers.

Conclusion

In his classic article, Mook (1983) introduced the distinction between testing whether effects "can" exist versus testing whether they "do" exist. When the primary goal of a research project is to test a prediction derived from theory, it is irrelevant whether the effect generalizes to other items or real-life settings. In such a context, it suffices to show that a certain effect "*can* be made to occur" (Mook, 1983, p. 385, italics added). Other research projects, however, are designed to explore a particular effect, to show that it plays an important role in a variety of settings, to quantify its size, and maybe even to "predict real-life behavior in the real world" (p. 381). The goal of these projects is to examine whether a certain effect "does" exist.

Most psychological research has focused on demonstrating that hypothesized effects "can" exist, using experimental materials and procedures specifically chosen to maximize the likelihood that the hypothesized effect occurs. Although this research has provided important insights, it is often unclear whether the observed effects actually play a role in everyday human cognition and behavior. Many of them do not replicate (Open Science Collaboration, 2015; Yong, 2012). Others are dependent on the specific stimuli used by the researchers (Bahník & Vranka, 2017; Westfall, Judd, & Kenny, 2015). If we want to know whether our effects *do* exist—that is, whether the psychological processes under investigation occur and play a role in real-world settings and the observed effects replicate even if a minor aspect of the study is changed—it will be necessary to randomly select items (i.e., stimuli, targets, confederates, schools, locations, settings) from a larger pool of possible items and to estimate LMEMs that include the appropriate by-item random effects in our data analyses.

LMEMs have numerous advantages. They allow us to analyze categorical and continuous predictors with the same data-analytic

[18] One alternative for dealing with small sample sizes and overparameterized/nonconverging models is to switch to Bayesian data analyses. Excellent texts on this topic have been published in recent years (e.g., Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2017; Stegmueller, 2013).

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

framework and with virtually the same script in the data analysis program, regardless of whether the predictors vary between or within "units" (i.e., subjects, groups, classrooms). In the case of categorical within-unit predictors, LMEMs perfectly reproduce the results of the repeated-measures ANOVA, as long as there are no missing values and the appropriate specifications are chosen (i.e., *F* approximation with Kenward-Roger *df*s, unstructured covariance matrix). Unlike repeated measures ANOVA, LMEMs can accommodate "unbalanced repeats" (e.g., different subjects produce a different number of data points for each level of the within-subject predictor).

LMEMs can handle issues that pose problems for other data-analytic strategies. For example, one can use LMEMs to analyze predictor variables with levels that are not equally spaced. LMEMs can easily handle data that contain multiple sources of nonindependence, regardless of whether this nonindependence is caused by subjects, targets, confederates, classrooms, settings, or locations. LMEMs are ideally suited to analyze change over time (or space) for subjects who are either independent or who are clustered by some higher-order unit (e.g., patients within therapists). LMEMs deal with missing data more effectively because unlike repeated measures ANOVAs, they do not eliminate all observations of a subject who has one or more missing values. It is a small step from LMEMs to "generalized linear-mixed effects models," which are needed when the outcome variables are categorical, ordinal, or counts.

Are standard statistical procedures such as ANOVA and regression analysis outdated? Definitely not, because there are still numerous psychological studies in which predictors vary between subjects and in which subjects is the only random variable. However, we often expose our subjects to multiple items (e.g., stimuli, confederates), examine how they change over time (e.g., therapy, learning), test them in multiple settings (e.g., locations, situations), and deal with subjects who influence each other (e.g., groups, classrooms). Given that these types of studies involve multiple sources of nonindependence, LMEMs will be increasingly present in the results sections of our scientific journals and undoubtedly belong in the standard tool kit for psychological researchers.

References

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

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,* 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005

Bahník, Š., & Vranka, M. A. (2017). If it's difficult to pronounce, it might not be risky: The effect of fluency on judgment of risk does not generalize to new stimuli. *Psychological Science, 28,* 427–436. http://dx.doi.org/10.1177/0956797616685770

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4,* 328. http://dx.doi.org/10.3389/fpsyg.2013.00328

AQ: 11

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,* 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015a). *Parsimonious mixed models*. Unpublished manuscript, University of Tübingen, Germany.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015b). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Brewer, P., & Venaik, S. (2014). The ecological fallacy in national culture research. *Organization Studies, 35,* 1063–1086. http://dx.doi.org/10.1177/0170840613517602

Clark, H. (1973). The language as fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359. http://dx.doi.org/10.1016/S0022-5371(73)80014-3

Demidenko, E. (2013). *Mixed models: Theory and applications with R*. New York, NY: Wiley.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12,* 121–138. http://dx.doi.org/10.1037/1082-989X.12.2.121

Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

Fullerton, A. S., & Xu, J. (2016). *Ordered regression models: Parallel, Partial, and non-parallel alternatives*. Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511790942

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). *Data analysis: A model comparison approach*. New York, NY: Routledge.

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,* 54–69. http://dx.doi.org/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,* 601–625. http://dx.doi.org/10.1146/annurev-psych-122414-033702

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

Kiernan, K., Tao, J., & Gibbs, P. (2012). *Tips and strategies for mixed modeling with SAS/STAT procedures (Paper 332–2012)*. Retrieved from http://support.sas.com/resources/papers/proceedings12/332-2012.pdf

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Kreft, I., & DeLeeuw, J. (1998). *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781849209366

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package lmerTest. R package version, 2–0. Retrieved from https://CRAN.R-project.org/package=lmerTest

AQ: 12

Liu, S., Rovine, M. J., & Molenaar, P. C. M. (2012). Selecting a linear mixed model for longitudinal data: Repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods, 22,* 15–30.

Loy, A., & Hofmann, H. (2014). HLMdiag: A suite of diagnostics for hierarchical linear models in R. *Journal of Statistical Software, 56,* 1–28.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13,* 203–229.

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. http://dx.doi.org/10.1016/j.jml.2017.01.001

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

McNeish, D., & Kelley, K. (2017). *Disentangling differences between fixed effects and mixed-effects models for clustered data*. Unpublished manuscript, University of North Carolina, Chapel Hill, NC.

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51,* 495–518. http://dx.doi.org/10.1080/00273171.2016.1167008

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387. http://dx.doi.org/10.1037/0003-066X.38.4.379

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., . . . Lewis, T. (2000). *A user's guide to MLwiN. Multilevel Models Project*. London, UK: Institute of Education, University of London.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E. J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22,* 304–321. http://dx.doi.org/10.1037/met0000057

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods, 13,* 150–170. http://dx.doi.org/10.1037/1082-989X.13.2.150

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution, 1,* 103–113. http://dx.doi.org/10.1111/j.2041-210X.2010.00012.x

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage Publishers.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science, 57,* 748–761. http://dx.doi.org/10.1111/ajps.12001

Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science, 10,* 390–399. http://dx.doi.org/10.1177/1745691614564879

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,* 2020–2045. http://dx.doi.org/10.1037/xge0000014

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59,* 1–23. http://dx.doi.org/10.18637/jss.v059.i10

Winter, B. (2016). *Convergence problem*. Retrieved from http://www.bodowinter.com/stuff/convergence_issues.pdf

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling. *Journal of Language Evolution, 1,* 7–18. http://dx.doi.org/10.1093/jole/lzv003

Wolsiefer, K., Westfall, J., & Judd, C. M. (2016). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods, 49,* 1210.

Yong, E. (2012). Replication studies: Bad copy. *Nature, 485,* 298–300. http://dx.doi.org/10.1038/485298a

Appendix A

AQ: 13

The Reshape Function in R

The reshape function in R is particularly well-suited to transform data files from one format into another, that is, from wide to long and from long to wide. The following R script transforms the data file in wide format that is shown in the top panel of Table 1 into the data file in long format that is shown in the bottom panel of Table 1:

```
## WIDE TO LONG
d_long<-reshape(d_wide,
    varying = c("like.lo1", "like.lo2", "like.hi1", "like.hi2"),
    v.names = "like",
    timevar = "prestige",
    times = c("1", "1", "2", "2"),
    new.row.names = 1:1000,
    direction = "long")
d_long <- d_long[order(d_long$subject.ID),]    # sort data file
d_long$id<-NULL    # eliminate unneeded variable
```

The following R script transforms the data file in long format that is shown in the bottom panel of Table 1 into the data file in wide format that is shown in the top panel of Table 1:

```
## LONG TO WIDE
d_long$index<-rep(c("lo1","lo2","hi1","hi2"),each = 1)    # index var
d_long$prestige<-NULL    # eliminate unneeded variable
d_wide <- reshape(d_long,
    timevar = "index",
    idvar = c("subject.ID", "gender", "age"),
    direction = "wide")
```

(*Appendices continue*)

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Appendix B

Within- and Between-Unit Associations

As mentioned in the main text, the centering of a continuous within-unit predictor affects the researchers' capacity to interpret its coefficient. When the predictor is centered around each unit's own mean (e.g., centered around each subject's own mean), its coefficient describes the within-unit association between the predictor and the outcome variable, and this association is likely to differ from the between-unit association.

The issue of within- versus between-unit associations may be particularly important in studies in which subjects are nested in higher-order units such as groups or classrooms. It may be that two variables are related at the group level (e.g., discussion groups that talk longer come up with more creative solutions) but not at the individual level (e.g., individuals who talk longer than their fellow discussion group members do not necessarily propose more creative solutions). It may even be that the same two variables are positively related at the individual level (e.g., individuals with a higher income tend to vote Republican) but negatively related at the aggregate level (e.g., states with a higher average income tend to be Democrat).

Enders and Tofighi (2007) describe a data-analytic procedure that allows researchers to examine the within-unit and the between-unit associations with the same model. It suffices to add the means of the higher-order unit (e.g., the group means) as a predictor to the model. The resulting LMEM contains two predictors, the group-mean centered predictor (describing the within-group association) and the group means (describing the between-groups associations), and researchers can examine whether each of them is reliably different from zero. Of course, the group-mean centered predictor varies within-groups (and therefore requires a by-group random slope) whereas the group means vary between-groups (and therefore do not require a by-group random slope). Enders and Tofighi (2007) also propose a test that allows researchers to examine whether the within-group association is reliably different from the between-groups association (p. 131).

If the study includes only a small number of higher-order units and a small number of lower-order units, the test of the within-group association and the test of the between-groups association may both be underpowered. If the researchers' hypothesis is not specifically related to one of the two types of associations but simply predicts an association between two variables in the subject population (which happens to be nested in higher-order units), they can proceed in the following way: (a) run a model with the group-mean centered predictor and the group means, (b) use Enders and Tofighi's (2007) procedure to determine whether the two types of associations are reliably different from each other, using a liberal significance level (e.g., $\alpha = .10$), and (c) if they are not, run a LMEM in which the outcome variable is regressed on the grand-mean centered predictor (including the appropriate by-group random effects). The resulting coefficient will be an amalgam of within- and between-groups association, but given that the two are not reliably different from each other, researchers would not be committing the ecological fallacy when they interpret this coefficient.

Later research on Enders and Tofighi's (2007) data analytic approach revealed that the estimate of the between-groups association tends to be biased toward the within-group association under certain circumstances (Lüdtke et al., 2008). For example, if the within-group association is small and the between-groups association is large, the coefficient associated with the group means will often underestimate the true population between-groups association. The coefficient will be biased especially when there are few subjects per higher-order unit and when the nonindependence introduced by the higher-order unit is relatively small. Lüdtke and colleagues suggested a data-analytic procedure that produces unbiased estimates (the so-called multilevel latent covariate approach), but an in-depth presentation of this approach goes beyond the scope of this article.

(*Appendices continue*)

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Appendix C

Is Item Always a Random Variable?

In some research projects, the question arises, which types of multiple measurements require by-item random effects. For example, do items from a scale require by-item effects? The answer is straightforward: When our materials can be thought of as a sample from a larger possible set and we want to generalize our findings to this set, then we want to include by-item random effects in the maximal random effects structure. Typical examples are faces from two or more social categories (e.g., subjects judge European American and African American faces) and words having certain characteristics (e.g., subjects react to emotional and neutral words). Said differently, when our goal is to generalize the findings from "our" items to the entire pool of items from which they were drawn, we need to include by-item random effects. Similarly, by-item random effects are required when a failure to replicate with new materials would undermine our hypothesis. Items of individual difference scales and test questions usually do not require by-item random effects. The same is true when items are included to activate a category and they do so indistinctly (i.e., there is effectively no variability due to items).

Sometimes, it is necessary to explicitly examine whether the effects of the predictor on the outcome varies considerably among the items included in the study. Wolsiefer, Westfall, and Judd (2016) recently examined different implicit attitudes measures for which it is generally assumed that stimulus words or pictures indistinctly activate the category or concept they are supposed to represent. The authors showed that contrary to the general assumption, the items caused nonindependence in the data. As a result, the test statistics from the traditional analyses were substantially inflated (by about 60%) compared with the LMEM analyses with the appropriate by-item random effects.

For most studies, it is quite clear whether a given variable is random or not. For some studies, however, the decision is less obvious. When the data file included observations from all the counties of one of the U.S. states and the goal is not to generalize to other U.S. counties, it is not entirely clear whether county is a random variable or a fixed variable (Snijders & Bosker, 2012). Although test questions are usually not considered a random variable, some researchers might argue that they should be. After all, the questions included in the test were selected from a larger pool of possible test questions and it would be worrisome if students had radically different scores on a different but similar test that included another subset of the large pool of possible test questions.

When the research question is entirely theory-driven, the goal of a study may be to show that a hypothesized effect can be made to occur, even if the stimulus material or circumstances are rather contrived. In other words, the goal may be to show that it is possible to produce a certain effect with items that were chosen for that particular purpose. In such research, the goal is not to generalize to similar items of the same kind, and the study's conclusion would not be invalidated if the results failed to replicate with different stimuli.

In a field study conducted on 12 different locations on campus, it is not entirely clear whether "location" is a random variable. One might argue that data have been gathered from all the levels of the variable that are of interest, a typical characteristic of fixed variables. The SAS User's Guide summarizes the situation appropriately: "One modeler's fixed effect is another modeler's random effect." It is acceptable to treat certain variables as fixed rather than random (i.e., to abstain from including certain random effects in the maximal random effects structure). However, such a choice should be justified in the manuscript, that is, authors should explain why they omitted to include the appropriate by-variable random slopes for variables that some of their colleagues might consider random.

(*Appendices continue*)

tapraid5/met-met/met-met/met99917/met2430d17z | xppws | S=1 | 9/9/17 | 3:41 | Art: 2016-0646

Appendix D

How to Compute the Number of Estimated Parameters

In the following paragraphs, we will provide a description of concrete studies and how to compute the number of parameters that are being estimated in a LMEM with the maximal random effects structure.

If there are four by-subject random effects (let's say one by-subject random intercept and three by-subject random slopes) and no by-item random effects, the LMEM will estimate a total of 10 parameters for the random effects structure alone: four parameters for the random effects and six parameters for all possible pairwise covariances of the random effects. LMEMs estimate the covariances of random effects belonging to the same unit (either subjects or items), but not the covariances of random effects belonging to different units. This is because it is impossible to compute the covariance between a by-subject random effect and a by-item random effect.

Consider a LMEM with four fixed effects, two by-subject random effects, and two by-item random effects (see Table 15). The model will estimate a total of 11 parameters: the four fixed effects, the four random effects, the covariance between the two by-subject random effects, the covariance between the two by-item random effects, and the residuals. Consider another LMEM with three predictors, all possible two- and three-way interactions, four by-subject random effects, and two by-item random effects (see Table 16). In total, the LMEM with the maximal random effects structure will estimate eight fixed effects, four by-subject random effects, six covariances between the by-subject random effects, two by-item random effects, one covariance between the by-item random effects, and the residuals . . . 22 parameters in total.

With more complex designs, the LMEM with the maximal random effects structure may attempt to estimate 100 parameters or more. Such a large number of parameters may cause problems in the estimation process. As mentioned above, maximum likelihood estimation involves an iterative process in which the computer progressively changes the parameter estimates with the goal to maximize the likelihood of having obtained the data at hand. With an excessive number of parameters and/or too many missing data and/or numerous random effects with near-zero variance, the model may fail to converge. In that case, the error message displayed by the data analysis program has to be taken seriously because the output that is printed below cannot be trusted.

Note that R assumes that users want to estimate all possible covariances. In all of the scripts presented in the tables of this article, R estimates all covariances among random effects that are listed in the same parenthesis.

Received May 24, 2016
Revision received June 12, 2017
Accepted June 22, 2017 ■