Random effects structure in mixed-effects models: Keep it maximal

UNDER REVIEW: Please do not cite

Dale J. Barr[a,*], Roger Levy[b], Christoph Scheepers[a], Harry J. Tily[c]

[a]*Institute of Neuroscience and Psychology*
University of Glasgow
58 Hillhead St.
Glasgow G12 8QB
United Kingdom
[b]*Department of Linguistics*
University of California at San Diego
La Jolla, CA 92093-0108
USA
[c]*Department of Brain and Cognitive Sciences*
Massachussetts Institute of Technology
Cambridge, MA 02139
USA

Abstract

Linear mixed effects models (LMEMs) are rapidly advancing as a candidate to replace ANOVA as a standard for inferential analyses in psycholinguistics and associated fields. However, because of the relative novelty of this approach, there are few clear standards regarding its correct use, as well as much uncertainty about whether it truly offers an advantage over traditional approaches. In this paper, we argue that many of the traditional standards in accounting for observational dependencies in the design also apply to the correct use of LMEMs. We argue that valid statistical inferences using LMEMs require maximal random-effects structures wherever possible—that is, including condition-specific random effects by subjects/items for every fixed effect of theoretical interest that is measured in more

*Corresponding author; Tel: +44 (0)141-330-1602; Fax: +44 (0)141-330-4606.
Email addresses: dale.barr@glasgow.ac.uk (Dale J. Barr), rlevy@ucsd.edu (Roger Levy), christoph.scheepers@glasgow.ac.uk (Christoph Scheepers), hjt@mit.edu (Harry J. Tily)

than one condition within subjects/items. We present both theoretical analysis and extensive Monte Carlo simulations in support of this argument. Failure to use maximal random-effects structures (e.g., the common practice of including only by-subject and by-item intercepts in a model's random effects specification) can lead to catastrophically inflated Type-I error rates. In contrast, using maximal random effect structures with LMEMs makes them the most flexible and widely-applicable approach available to date.

Keywords: linear mixed-effects models, generalization, statistics, Monte Carlo simulation

---

The notion of *independent evidence* plays no less important a role in the assessment of scientific hypotheses than it does in everyday reasoning. Consider a pet-food manufacturer determining which of two new gourmet cat-food recipes to bring to market. The manufacturer has every interest in choosing the recipe that the average cat will eat the most of. Thus every day for a month (twenty-eight days) their expert, Dr. Nyan, feeds one recipe to a cat in the morning and the other recipe to a cat in the evening, counterbalancing which recipe is fed when and carefully measuring how much was eaten at each meal. At the end of the month Dr. Nyan calculates that recipes 1 and 2 were consumed to the tune of 92.9±5.6 and 107.2±6.1 (means ± SDs) grams per meal respectively. How confident can we be that recipe 2 is the better choice to bring to market? Without further information you might hazard the guess "somewhat confident", considering that one of the first statistical hypothesis tests typically taught, the unpaired t-test, gives $p = 0.09$ against the null hypothesis that choice of recipe does not matter. But now we tell you that only seven cats participated in this test, one for each day of the week. How does this change your confidence in the superiority of recipe 2?

Let us first take a moment to consider precisely what it is about this new information that might drive us to change our analysis. The unpaired t-test is based on the assumption that all observations are *conditionally independent* of one another given the true underlying means of the two populations—here, the average amount a cat would consume of each recipe in a single meal. Since no two cats are likely to have identical dietary proclivities, multiple measurements from the same cat would violate this assumption. The correct characterization becomes that all observations are conditionally independent of one another given (a) the true palatibility effect of recipe 1 versus recipe 2, together with (b) the dietary proclivities of each cat. This weaker conditional independence is a double-edged sword. On the one hand, it means that we have tested effectively fewer individu-

als than our 56 raw data points suggest, and this should weaken our confidence in generalizing the superiority of recipe 2 to the entire cat population. On the other hand, the fact that we have made multiple measurements for each cat holds out the prospect of factoring out each cat's idiosyncratic dietary proclivities as part of the analysis, and thereby improving the signal-to-noise ratio for inferences regarding each recipe's overall appeal. How we specify these idiosyncrasies can dramatically affect our inferences. For example, we know that some cats have higher metabolisms and will tend to eat more at every meal than other cats. But we also know that each creature has its own palate, and even if the recipes were of similar overall quality, a given cat might happen to like one recipe more than the other. Indeed, accounting for idiosyncratic recipe preferences for each cat might lead to even weaker evidence for the superiority of recipe 2.

Situations such as these, where individual observations cluster together via association with a smaller set of entities, are ubiquitous in the language sciences—though of course the clusters are typically human participants and different types of linguistic items, rather than cats and cat-food recipes. Similar clustered-observation situations arise in other sciences, such as agriculture (plots in a field) and sociology (students in classrooms in schools in school-districts); hence accounting for the RANDOM EFFECTS of these entities has been an important part of the workhorse statistical analysis technique, the ANALYSIS OF VARIANCE, under the name MIXED-MODEL ANOVA, since the first half of the twentieth century (Fisher, 1925; Scheffe, 1959). In experimental psychology, the prevailing standard for a long time has been to assume that individual participants may have idiosyncratic sensitivities to any experimental condition that may have an overall effect, so detecting a "fixed effect" of some manipulation must be done under the assumption of corresponding participant random effects for that manipulation as well. In our pet-food example, if there is a true effect of recipe—that is, if on average a new, previously unstudied cat will on average eat more of recipe 2 than of recipe 1—it should be detectable above and beyond the noise introduced by cat-specific recipe preferences. Technically speaking, the fixed effect is tested against an error term that captures the variability of the effect across individuals.

Standard practices for data-analysis in psycholinguistics fundamentally changed, however, after Clark (1973). In a nutshell, Clark (1973) argued that linguistic materials, just like experimental participants, have idiosyncrasies that need to be accounted for. Because in a typical psycholinguistic experiment, there are multiple observations for the same item (e.g., a given word or sentence), these idiosyncrasies break the conditional independence assumptions underlying mixed-model ANOVA which treats experimental participant as the only random effect. Clark

proposed the min-F' statistic as a conservative approximation to an F-ratio whose distributional assumptions are satisfied even under what in contemporary parlance is called CROSSED random effects of participant and item (Baayen et al., 2008). Clark's paper helped drive the field toward a standard demanding evidence that experimental results generalized beyond the specific linguistic materials it used—in other words, the so-called by-subjects F_1 mixed-model ANOVA was not enough. There was even a time where reporting of the min-F' statistic was made a standard for publication in the *Journal of Memory and Language*. However, acknowledging the widespread belief that min-F' is unduly conservative (but see Forster & Dickinson, 1976), significance of min-F' was never made a requirement for acceptance of a publication. Instead, the 'normal' convention continued to be that a result is considered likely to generalize if it passes $p < 0.05$ significance in both by-subjects (F_1) and by-items (F_2) ANOVAs. In the literature this criterion is called $F_1 \times F_2$ (e.g., Forster & Dickinson, 1976), which in this paper we use to denote the larger (less significant) of the two p values derived from F_1 and F_2 analyses.

Linear Mixed-Effects Models (LMEMs)

Since Clark (1973), the biggest change in data analysis practices has been the introduction of methods for simultaneously modeling crossed participant and item effects in a single analysis, in what is variously called "hierarchical regression", "multi-level regression", or simply "mixed-effects models" (Baayen et al., 2008; Gelman & Hill, 2007; Goldstein, 1995; Locker et al., 2007; Pinheiro & Bates, 2000; Quené & van den Bergh, 2008; Snijders & Bosker, 1999).[1] In this paper we refer to models of this class as *mixed-effects models*; when fixed effects, random effects, and trial-level noise contribute *linearly* to the dependent variable, it is a *linear mixed-effects model* (LMEM).

In addition to their ability to handle crossed random effects, mixed-effects models enjoy a number of in-principle advantages over ANOVA, as others have noted:

- they do not assume balanced datasets; hence, they handle missing data and

[1] We should emphasize here that despite the "mixed-effects models" nomenclature, traditional ANOVAs used in psycholinguistics have always used "mixed effects"; what is new is the explicit, simultaneous estimation of both fixed-effects and random-effects components of such a model. This permits correct treatment of imbalanced data and, as so clearly indicated by the title of Baayen et al. (2008), allows extension to *crossing* of two or more types of random effects with a set of fixed effects in a single analysis.

naturalistic datasets (e.g., language corpora) more gracefully and flexibly than ANOVAs;

- they require specification of a complete generative process hypothesized to underlie observed data, and the model components can be explicitly estimated and inspected;
- they allow for improved inferences about random effects themselves, which can be useful for purposes such as individual-differences studies;
- Generalized LMEMs can faithfully model non-normally distributed and even categorical data;
- Because imbalance is not problematic, a wide variety of categorical and continuous predictors, both those of critical theoretical interest and controls, can be included together in a single analysis.

Over recent years, LMEMs have thus enjoyed widespread adoption and possess considerable momentum as a candidate to replace ANOVA as a standard in language research. What remains less widespread, however, is consensus regarding what standards should apply to mixed-effects analysis. In this paper, we focus on one issue in particular: when the goal of an analysis is to make reliable inferences about one or more "fixed effects", what random-effects structure should one use? Based on theoretical analysis and Monte Carlo simulation, we argue the following:

1. Implicit choices regarding random-effect structures existed for traditional mixed-model ANOVAs just as they exist today for LMEMs;
2. With mixed-model ANOVAs, the standard has been to use "maximal" random-effect structures;
3. Insofar as we as a field think this standard was appropriate in the past, it is appropriate today and LMEMs should, whenever possible, also use "maximal" random-effect structures;
4. Failure to include maximal random-effect structures in LMEMs (when random effects are present in the underlying populations) inflates Type I error rates;
5. LMEMs with random intercepts only have catastrophically high Type I error rates, regardless of how p-values are computed from them (see also Roland, 2009; Jaeger, 2011; and Schielzeth & Forstmeier, 2009);

6. Forward stepwise "model fitting" approaches for determining random effects also show unacceptably high Type I error rates in many cases;
7. Contrary to some warnings in the literature (Pinheiro & Bates, 2000), likelihood-ratio tests for fixed effects in LMEMs show minimal Type I error inflation for psycholinguistic datasets (see Baayen et al., 2008, Footnote 1, for a similar suggestion);
8. The $F_1 \times F_2$ criterion leads to increased Type I error rates (i) the more the effects vary across subjects and items in the underlying populations and (ii) the larger the sample sizes are (see also Clark, 1973; Forster & Dickinson, 1976);
9. Min-F' is conservative in between-items designs when the item variance is low, and is conservative overall for within-items designs, especially so when the treatment-by-subject and/or treatment-by-item variances are low (see also Forster & Dickinson, 1976); in contrast, maximal LMEMs show no such conservativity.

Specifying random effects: You already know how

The *Journal of Feline Gastronomy* has just received a submission reporting that the feline palate prefers tuna to liver, and as journal editor you must decide whether to send it out for review. The authors report a highly significant effect of recipe type ($p < .0001$), stating that they used "a mixed effects model with random effects for cats and recipes". Are you in a position to evaluate the generality of the findings? Given that LMEMs can implement nearly any of the standard parametric tests—from a one-sample test to a multi-factor mixed-model ANOVA—the answer can only be no. Indeed, whether LMEMs produce valid inferences depends critically on *how* they are used. What you need to know in addition is the *random effects structure* of the model, because this is what the assessment of the treatment effects is based on. In other words, you need to know which treatment effects are assumed to vary across which sampled units, and how they are assumed to vary. As we will see, whether one is specifying a random effects structure for LMEMs or choosing an analysis from among the traditional options, the same considerations come into play. So, if you are scrupulous about choosing the "right" statistical technique, then you should be equally scrupulous about using the "right" random effects structure in LMEMs.

In fact, knowing how to choose the right test already puts you in a position to specify the correct random effects structure for LMEMs. In this section, we attempt to distill the implicit standards already in place by walking through a hypothetical example and discussing the various models that could be applied,

their underlying assumptions, and how these assumptions relate to more traditional analyses. In our hypothetical experiment, subjects perform a lexical decision to a string of letters. Each subject is exposed to two types of words, forming condition A and condition B of the experiment. The words in one condition differ from those in the other condition on some intrinsic dimension (e.g., frequency, imageability, etc.), comprising a word-type manipulation that is within-subjects and between-items. The question is whether reaction times are systematically different between condition A and condition B. For expository purposes, we use a "toy" dataset with hypothetical data from four subjects and four items, yielding two observations per treatment condition per participant. The observed data are plotted alongside predictions from the various models we will be considering in the panels of Figure 1. Because we are using simulated data, all of the parameters of the population are known, as well as the "true" subject-specific and item-specific effects for the sampled data. In practice, researchers do not know these values and can only estimate them from the data; however, using known values for hypothetical data can aid in understanding how clustering in the population maps onto clustering in the sample.

The general pattern for the observed data points suggests that items of type B (I3 and I4) are responded to faster than items of type A (I1 and I2). This suggests a simple (but clearly inappropriate) model for these data that relates response Y_{si} for subject s and item i to a baseline level β_0, a treatment effect β_1, and observation-level error e_{si}:

$$Y_{si} = \beta_0 + \beta_1 X_i + e_{si} \tag{1}$$

where X_i is a predictor variable taking on the value of 0 or 1 depending on whether item i is of type A or B respectively.[2] In the population, participants respond to items of type B 40 ms faster than items of type A. Under this first model, we assume that each of the 16 observations provides the same evidence for or against the treatment effect regardless of whether or not any other observations have already been taken into account. Performing a (clearly inappropriate) unpaired t-test

[2] In this example, we use a "contrast coding" scheme (0 or 1) for the predictor variable. Alternatively, the models in this section could be expressed in the style more common to traditional ANOVA pedagogy, where fixed and random effects represent deviations from a grand mean. This model can be fit by using "deviation coding" for the predictor variable (-.5 and .5 instead of 0 and 1). For higher-order designs, contrast coding and deviation coding schemes will lead to different interpretations for lower-order effects (simple effects for contrast coding and main effects for deviation coding).

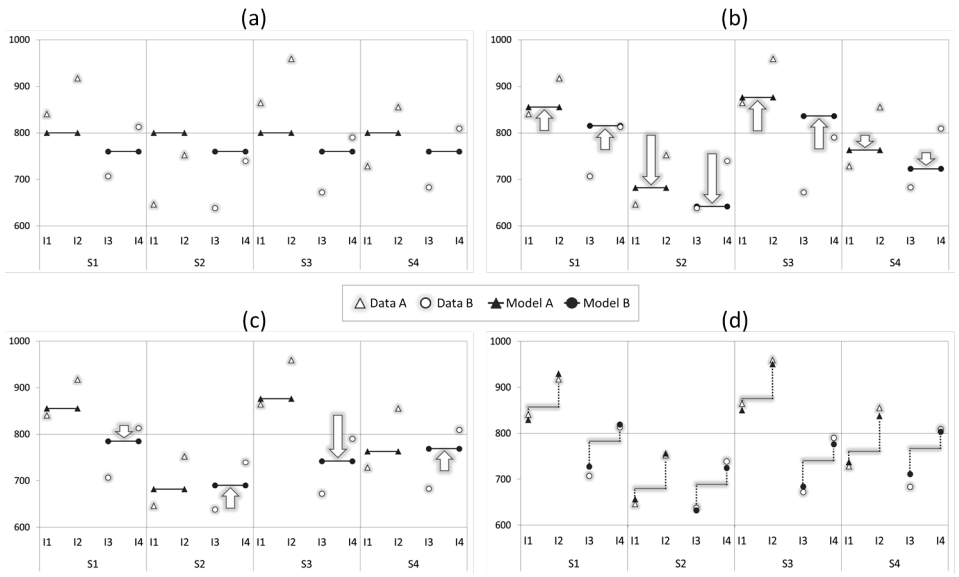

Figure 1: Example RT data (open symbols) and model predictions (filled symbols) for a hypothetical lexical decision experiment with two within-subject/between-item conditions, A (triangles) and B (circles), including four subjects (S1-S4) and four items (I1-I4). Panel (a) illustrates a model with no random effects, considering only the baseline average RT (response to word type A) and treatment effect; panel (b) adds random subject intercepts to the model; panel (c) adds by-subject random slopes; and panel (d) illustrates the additional inclusion of by-item random intercepts. Panel (d) represents the maximal random-effects structure justified for this design; any remaining discrepancies between observed data and model estimates are due to trial-level measurement error (e_{si}).

on these data would implicitly assume this generative model.

Model (1) is not a mixed-effects model because we have not defined any sources of clustering in our data; all observations are conditionally independent given a choice of intercept, treatment effect, and noise level. But experience tells us that different subjects are likely to have different overall response latencies, breaking conditional independence between trials for a given subject. We can expand our model to account for this by including a new offset term S_{0s}, the deviation from β_0 for subject s. The expanded model is now

$$Y_{si} = \beta_0 + S_{0s} + \beta_1 X_i + e_{si}. \tag{2}$$

These offsets increase the model's expressivity by allowing predictions for each subject to shift upward or downward by a fixed amount (Figure 1b). Our use of Latin letters for this term is a reminder that S_{0s} is a special type of effect which is

different from the βs—indeed, we now have a "mixed-effects" model: parameters β_0 and β_1 are *fixed effects* (effects that are assumed to be constant from one experiment to another), while the specific composition of subject levels for a given experiment is assumed to be a random subset of the levels in the underlying populations (another instantiation of the same experiment would have a different composition of subjects, and therefore different realizations of the S_{0s} effects). The S_{0s} effects are therefore *random effects.* Our primary goal is to produce a model which will generalize to the *population* from which these subjects are randomly drawn, rather than describing the specific S_{0s} values for this sample. Therefore, instead of estimating the individual S_{0s} effects, the model-fitting algorithm estimates the population distribution from which the S_{0s} effects were drawn. This requires assumptions about this distribution; commonly, these assumptions are that it is normal, with a mean of 0 and a variance of ${\tau_{00}}^2$; here ${\tau_{00}}^2$ is a *random effect parameter*, and is denoted by a Greek symbol because, like the βs, it refers to the population and not to the sample.

Fitting Model (2) is analogous to analyzing the raw, unaggregated response data using a repeated-measures ANOVA with $SS_{subjects}$ subtracted from the residual SS_{error} term. Although Model (2) is clearly preferable to Model (1), it does not capture all the possible by-subject dependencies in the sample; experience also tells us that subjects often vary not only in their overall response latencies *but also in the nature of their response to word type*. In the present hypothetical case, Subject 3 shows a total effect of 134 ms, which is 94 ms larger than the average effect in the population of 40 ms. We have multiple observations per combination of subject and word type, so this variability in the population will also create clustering in the sample. The S_{0s} do not capture this variability; they are *random intercepts* that allow subjects to vary around β_0. What we need in addition are *random slopes* to allow subjects to vary with respect to β_1, our treatment effect. We introduce random slope term S_{1s}, which yields

$$Y_{si} = \beta_0 + S_{0s} + (\beta_1 + S_{1s})X_i + e_{si}. \tag{3}$$

This is now a mixed-effects model with by-subject *random intercepts and slopes*. Note that the inclusion of the by-subject random slope causes the predictions for condition B to shift by a fixed amount for each subject (Figure 1c), improving predictions for words of type B. The slope offset S_{1s} captures how much Subject s's effect deviates from the population treatment effect β_1. Again, we do not want to our analysis to commit to particular S_{1s} effects, and so, rather than estimating these values directly, we estimate ${\tau_{11}}^2$, the by-subject variance in treatment ef-

fect. But note that now we have two random effects for each subject s, and these two effects can be correlated. For example, subjects who do not read carefully might not only respond faster than the typical subject (and have a negative S_{0s}), but might also show less sensitivity to the word type manipulation (and have a more positive S_{1s}). Indeed, such a negative correlation is suggested in our hypothetical data (Figure 1): S1 and S3 are slow responders who show clear treatment effects, whereas S2 and S4 are fast responders who are hardly susceptible to the word type manipulation. We therefore cannot treat these effects as coming from independent univariate distributions, but instead need to estimate the bivariate distribution from which the S_{0s} effects and S_{1s} effects are drawn. The latter can be described by three parameters: ${\tau_{00}}^2$ (random intercept variance), ${\tau_{11}}^2$ (random slope variance), and ${\tau_{01}}^2$ (the intercept/slope covariance). This distribution is assumed to be bivariate normal with a mean of $(0, 0)$.

Both Models (2) and (3) are instances of what is traditionally analyzed using "mixed-model ANOVA" (e.g., Scheffe, 1959, chapter 8). By long-standing convention in our field, however, the classic "by-subjects ANOVA" (and analogously "by-items ANOVA" when items are treated as the random effect) is understood to mean Model (3), the relevant F-statistic for which is $F_1 = \frac{MS_T}{MS_{TxS}}$, where MS_T is the treatment mean square and $\mathrm{MS}_{\mathrm{TxS}}$ is the treatment-by-subject mean square. This convention presumably derives from the widespread recognition that subjects (and items) usually *do* vary idiosyncratically not only in their global mean responses but also in their sensitivity to the experimental treatment.

Model (3) is also the model that was criticized by Clark (1973), since the repetition of words across observations is a source of non-independence not accounted for by Model (3). To complete the model, we also need to incorporate item variability with the random effect I_{0i}, yielding

$$Y_{si} = \beta_0 + S_{0s} + I_{0i} + (\beta_1 + S_{1s})X_i + e_{si}. \tag{4}$$

This is a mixed-effect model with by-subject random intercepts and slopes and by-item random intercepts. Note that the inclusion of by-item random intercepts improves the predictions from the model, with predictions for a given item shifting by a consistent amount across all subjects (Figure 1d). Rather than committing to specific I_{0i} values, we assume that the I_{0i} effects are drawn from a normal distribution with a mean of zero and variance ${\omega_{00}}^2$. We also assume that ${\omega_{00}}^2$ is independent from the τ parameters defining the by-subject variance components.

This analysis has a direct analogue to min-F', which tests MS_T against a denominator term consisting of the sum of MS_{TxS} and MS_I, the mean squares for

the random treatment-by-subject interaction and the random main effect of items. It is, however, different from the practice of performing $F_1 \times F_2$ and rejecting the null hypothesis if $p < .05$ for both Fs. The reason is that MS_T (the numerator for both F_1 and F_2) reflects not only the treatment effect, but also treatment-by-subject variability (τ_{11} as well as by-item variability (${\omega_{00}}^2$). The denominator of F_1 controls for treatment-by-subject variability but not item variability; similarly, the denominator of F_2 controls for item variability but not treatment-by-subject variability. Thus, finding that F_1 is significant implies that $\beta_1 \neq 0$ or $\omega_{00}^2 \neq 0$, or both; likewise, finding that F_2 is significant implies that $\beta_1 \neq 0$ or $\tau_{11}^2 \neq 0$, or both. Since ${\omega_{00}}^2$ and ${\tau_{11}}^2$ can be nonzero while $\beta_1 = 0$, $F_1 \times F_2$ tests will inflate the Type I error rate (Clark, 1973). Thus, in terms of controlling Type I error rate, the mixed-effects modeling approach and the min-F' approach are, at least theoretically, superior to separate by-subject and by-item tests.

At this point, we might wish to go further and consider other models. For instance, we have considered a by-subject random slope; for consistency, why don't we also consider a model with a by-item random slope, I_{1i}? A little reflection reveals that a by-item random slope does not make sense for this design. Words are nested within word types—no word can be both type A and type B—so it is not sensible to ask whether words vary in their sensitivity to word type. No sample from this experiment could possibly give us the information needed to estimate random slope variance and random slope/intercept covariance parameters for such a model. A model like this is *unidentifiable* for the data it is applied to: there are (infinitely) many different values we could choose for its parameters which would describe the data equally well.[3] Experiments with a within-item manipulation, such as a priming experiment in which target words are held constant across conditions but the prime word is varied, would call for by-item random slopes, but not the current experiment.

The above point also extends to designs where one independent variable is manipulated *within-* and another variable *between-* subjects (respectively items). In case of between-subject manipulations, the levels of the subject variable are nested within the levels of the experimental treatment variable (i.e. each subject belongs to one and only one of the experimental treatment groups), meaning that subject and treatment cannot interact—a model with a by-subject random

[3]Technically, by-item random slopes for a between-item design can be used to capture heteroscedasticity across conditions, but this is typically a minor concern in comparison with the issues focused on in this paper (see, e.g., discussion in Gelman & Hill, 2007).

slope term would be unidentifiable. In general, within-unit treatments require both the by-unit intercepts and slopes in the random effects specification, whereas between-unit treatments require (and logically permit) only the by-unit random intercepts.

It is important to note that identifiability is a property of the model given a certain dataset. The model with by-item random slopes is unidentifiable for any possible dataset because it tries to model a source of variation that could not logically exist in the population. However, there are also situations where a model is unidentifiable because there is insufficient data to estimate its parameters. For instance, we might decide it was important to try to estimate variability corresponding to the different ways that subjects might respond to a given word (a subject-by-item random intercept). But to form a cluster in the sample, it is necessary to have more than one observation for a given unit; otherwise, the clustering effect cannot be distinguished from residual error. If we only elicit one observation per subject/item combination, we are unable to estimate this source of variability, and the model containing this random effect becomes unidentifiable. Had we used a design with repeated exposures to the same items for a given subject, the same model would be identifiable, and in fact we would need to include that term to avoid violating the conditional independence of our observations given subject and item effects.

This discussion indicates that Model (4) has the maximal random effects structure justified by our experimental design. A model with maximal random effects structure optimizes generalization of the findings to new subjects and new items. Models that lack random effects contained in the maximal model, such as Models (1)-(3), are *misspecified*—the model specification is not expressive enough to include the true generative process underlying the data. This type of misspecification is bad because conditional independence between observations within a given cluster is not achieved. This will typically underestimate the standard errors of fixed-effects and, consequently, inflate the Type I error rates for fixed effects.

A final model that we have not considered yet is one that includes by-subject and by-item random intercepts only (RI-only model).

$$Y_{si} = \beta_0 + S_{0s} + I_{0i} + \beta_1 X_i + e_{si} \tag{5}$$

One might be tempted to consider a model such as (5) if, for example, the treatment-by-subject variability (${\tau_{11}}^2$) is of little theoretical interest (Locker et al., 2007), or if a comparison of models (4) and (5) suggests that the simpler model is warranted (Baayen et al., 2008). However, this would be a mistake. Even if ${\tau_{11}}^2$ is not of

Table 1: Summary of models considered and associated `lmer` syntax.

No.	Model	`lmer` model syntax
(1)	$Y_{si} = \beta_0 + \beta_1 X_i + e_{si}$	n/a (not a mixed-effects model)
(2)	$Y_{si} = \beta_0 + S_{0s} + \beta_1 X_i + e_{si}$	`Y~C+(1|Subject)`
(3)	$Y_{si} = \beta_0 + S_{0s} + (\beta_1 + S_{1s})X_i + e_{si}$	`Y~C+(1+C|Subject)`
(4)	$Y_{si} = \beta_0 + S_{0s} + I_{0i} + (\beta_1 + S_{1s})X_i + e_{si}$	`Y~C+(1+C|Subject)+(1|Item)`
(5)	$Y_{si} = \beta_0 + S_{0s} + I_{0i} + \beta_1 X_i + e_{si}$	`Y~C+(1|Subject)+(1|Item)`

theoretical interest, it is nonetheless a source of clustering in the data, and ignoring it is a model misspecification that breaks conditional independence between observations within a subject or item cluster. Unlike for the other models we have considered up to this point, there is no precedent for this kind of RI-only analysis in the mixed-model ANOVA literature. Still, it is possible to construe it as tantamount to a modified F'_{min} statistic, in which MS_{TxS} has been removed from the denominator and replaced with MS_S. But this would be clearly problematic because the numerator MS_T increases as a function of not only the treatment effect, but also the treatment-by-subject variability (${\tau_{11}}^2$). This implies that RI-only models should have an unacceptably high Type I error rate, increasing as a function of ${\tau_{11}}^2$, the by-subject random slope parameter. Indeed, this Type I inflation has been already noted by Roland (2009) and Schielzeth & Forstmeier (2009). The latter found a high (up to .41) Type I error rate for RI-only models on within-subject data (although they were not concerned, as we are, with the problem of simultaneous generalization).

The mixed-effects models considered in this section are presented in Table 1. We give their expression in the syntax of `lmer`, a widely used mixed-effects fitting method for R (Bates et al., 2011). To summarize, when specifying random effects, one must be guided by (1) the sources of clustering that exist in the target subject and item populations, and (2) by whether this clustering in the population will also exist in the sample. The general principle is that a by-subject (or by-item) random intercept is needed whenever there is more than one observation per subject (or item or subject/item combination), and a random slope is needed for any effect where there is more than one observation for each unique combination of subject and treatment level (or item and treatment level, or subject/item combination and treatment level). Models are unidentifiable when they include random effects that are logically impossible or that cannot be estimated from the data in

principle. Models are misspecified when they fail to include random effects that create dependencies in the sample. Although this section has only dealt with a simple single-factor design, these principles extend in a straightforward manner to higher-order designs, which we consider further in the General Discussion.

Against a model selection approach to random effects specification

The previous section implies a strong theoretical commitment to maximal random effect structures, i.e. models that include all random effect terms that are logically possible and statistically determinable for given experimental design. This ensures that we account for all potentially relevant dependencies *regardless* of whether conclusive evidence for these dependencies can be found in one's data.

An alternative way of specifying random effects structures is model selection. The basic idea behind this approach is to let the data "speak for themselves" as to whether certain random effect terms should be included in the model or not. That is, on the same data set, one compares the fit of a model with and without certain random effect terms (e.g. Model 4 versus Model 5 in the previous section) using goodness of fit criteria that take into account both the accuracy and the complexity of the model. Here, *accuracy* refers to how much variance is explained by the model and *complexity* to how many predictors (or parameters) are included in the model. The goal is to find an optimal compromise between the two: we want accurate, but simple models. This "optimal" random effect structure is then used when carrying out hypothesis tests on the fixed effects of interest. Indeed, a number of recent publications using LMEMs are based on such an approach (see following section). However, there are a number of theoretical and practical caveats which, by more than just coincidence, are related to those in step-wise multiple regression and other applications involving data-driven model optimization.

First, the order in which effects are entered into the model can have an influence on the outcome. Broadly speaking, there are two ways in which model building and comparison can proceed, known respectively as *forward selection* and *backward selection* methods. In forward selection, one starts with a simple model (e.g. Model 1 in the previous section) and makes it successively more complex (by adding random intercept and slope terms) until just before it becomes "unjustifiably complicated." Backward selection proceeds in the opposite direction, starting with a maximally complex model (e.g. Model 4 in the previous section) and making it successively simpler until just before it becomes "unbearably inaccurate." Importantly, neither of these heuristics performs an exhaustive exploration of the entire range of possible models, but rather terminates at some point along the way of making the model more (respectively less) complex. This can

introduce bias because one can easily get stuck in a *local optimum* (particularly with more complex factorial designs): the procedure terminates even though there might be better models further down the line of the model complexity/simplicity hierarchy.

It is also important to note that experimental designs are typically optimized for the detection of fixed effects, not of random effects. Simplicity-based model selection will therefore not only (correctly) reject random effects that do not exist, but also (incorrectly) reject random effects for which there is just insufficient power. This problem is exacerbated for datasets with few subjects and/or items, since detecting random effects is harder the fewer clusters are present. One can imagine the logical extreme of a study with only a single subject: a model-selection approach would invariably reject all by-subject random effects and thus merrily conclude that any fixed effect was likely to generalize over subjects; by contrast, the maximal random-effects model would correctly refuse to answer the ill-posed question of generalization over subjects in this case.

These arguments against model selection are perhaps less severe in the context of large-scale investigations involving plenty of data per design cell (e.g. research on language corpora, provided there are no sparse data problems). Nevertheless, in terms of generalization, theoretically motivated, maximal random effect structures should be superior to empirically selected ones.

Modeling of random effects in the current psycholinguistic literature

The introduction of LMEMs and their early application to psycholinguistic data by Baayen et al. (2008) has had a major influence on analysis techniques used in peer-reviewed publications. At the time of writing, Google scholar reports 408 citations to Baayen, Davidson and Bates. In a informal survey of the 150 articles appearing in the *Journal of Memory and Language* since (from 59(4) to 64(3)), we found that 20 (13%) reported analyses using an LMEM of some kind. However, these papers differ substantially in both the type of models used and the information reported about them. In particular, researchers differ in whether they include random slopes or only random intercepts in their models. Of the 20 JML articles identified, six give no information about the random effects structure, and a further six specify that they use random intercepts only, without theoretical or empirical justification. A further five papers employ model selection, four forward and only one backward. The final three papers employ a maximal random effects structure including intercept and slope terms where appropriate.

This survey highlights two important points. First, there appears to be no standard in reporting the modeling procedure, and authors vary dramatically in the

amount of detail they provide. Second, at least 30% of the papers using LMEMs, and perhaps as many as 60%, do not include random slopes, i.e. they tacitly assume that individual subjects and items are affected by the experimental manipulations in exactly the same way. As discussed earlier, this is a departure even from the standard use of ANOVA in psycholinguistics.

The present study

How do current uses of LMEMs compare to more traditional methods such as min-F' and $F_1 \times F_2$)? The next section of this paper tests a wide variety of commonly used analysis methods for datasets typically collected in psycholinguistic experiments, both in terms of whether resulting significance levels can be trusted—i.e., whether a p-value is *conservative* (less than α), *nominal* (equal to α), or *anticonservative* (greater than α)—and the *power* of each method in detecting effects that are actually present in the populations.

Ideally, we would compare the different analysis techniques by applying them to a large selection of real data sets. Unfortunately, in real experiments the true generative process behind the data is unknown, meaning that we cannot tell whether effects in the population exist—or how big those effects are—without relying on one of the analysis techniques we actually want to test. Moreover, even if we knew which effects were real, we would need far more datasets than are readily available to reliably estimate the nominality and power of a given method.

We therefore take an alternative approach of using Monte Carlo methods to generate data from simulated experiments. This allows us to specify the underlying sampling distributions per simulation, and thus to have veridical knowledge of the presence or absence of an effect of interest, as well as all other properties of the experiment (number of subjects, items and trials, and the amount of variability introduced at each level). Such a Monte Carlo procedure is standard for this type of problem (e.g., Davenport & Webster, 1973; Forster & Dickinson, 1976; Quené & van den Bergh, 2004; Santa et al., 1979; Schielzeth & Forstmeier, 2009; Wickens & Keppel, 1983), and guarantees that as the number of samples increases, the obtained p-value distribution becomes arbitrarily close to the true p-value distribution for datasets generated by the sampling model.

Moreover, the simulations assume a "best-case scenario" in which all the distributional assumptions of the model class (in particular normal distribution of random effects and trial-level error, and homoscedasticity of trial-level error and between-items random intercept variance) are satisfied. Although the approach leaves open for future research many difficult questions regarding departures of

realistic psycholinguistic data from these assumptions, it allows us great flexibility in analyzing the behavior of each analytic method as the population and experimental design vary. We hence proceed to the systematic investigation of traditional ANOVA, min-F', and several types of LMEMs as datasets vary in many crucial respects including between- versus within-items, different numbers of items, and different random-effect sizes and covariances.

Method

For simplicity, all datasets included a continuous response variable and had only a single two-level treatment factor, which was always within subjects, and either within or between items. When it was within, each "subject" was assigned to one of two counterbalancing "presentation" lists, with half of the subjects assigned to each list. We assumed no list effect; that is, the particular configuration of "items" within a list did not have any unique effect over and above the item effects for that list. We also assumed no order effects, nor any effects of practice or fatigue. All experiments had 24 subjects, but we ran simulations with both 12 or 24 items to explore the effect of number of random-effect clusters on fixed-effects inference.[4]

Within-item data sets were generated from the following sampling model: $Y_{si} = \beta_0 + S_{0s} + I_{0i} + (\beta_1 + S_{1s} + I_{1i})X_{si} + e_{si}$. with all variables defined as in the tutorial section above, except that we used deviation coding for X_{si} (-.5, .5) rather than contrast coding. Random effects S_{0s} and S_{1s} were drawn from a bivariate normal distribution with means $\mu_S = < 0, 0 >$ and variance-covariance matrix $\mathrm{T} = \begin{pmatrix} {\tau_{00}}^2 & \rho_I \tau_{00} \tau_{11} \\ \rho_I \tau_{00} \tau_{11} & {\tau_{11}}^2 \end{pmatrix}$. Likewise, I_{0i} and I_{1i} were also drawn from a separate bivariate normal distribution with $\mu_I =< 0, 0 >$ and variance-covariance matrix $\Omega = \begin{pmatrix} {\omega_{00}}^2 & \rho_S \omega_{00} \omega_{11} \\ \rho_S \omega_{00} \omega_{11} & {\omega_{11}}^2 \end{pmatrix}$. The residual errors e_{si} were drawn from a normal distribution with a mean of 0 and variance σ^2. For between-item designs, the I_{1i} effects (by-item random slopes) were simply ignored and thus did not contribute to the response variable.

We investigated the performance of various analyses over a range of population parameter values (Table 2). To generate each simulated dataset, we first

[4]Having only six items per condition, such as in the 12-item case, is not uncommon in psycholinguistic research, where is often difficult to come up with larger numbers of suitably controlled items.

Table 2: Ranges for the population parameters; $\sim U(min, max)$ means the parameter was sampled from a uniform distribution with range $[min, max]$.

Parameter	Description	Value
β_0	grand-average intercept	$\sim U(-3, 3)$
β_1	grand-average slope	0 (H_0 true) or .8 (H_1 true)
${\tau_{00}}^2$	by-subject variance of S_{0s}	$\sim U(0, 3)$
${\tau_{11}}^2$	by-subject variance of S_{1s}	$\sim U(0, 3)$
ρ_S	correlation between (S_{0s}, S_{1s}) pairs	$\sim U(-.8, .8)$
${\omega_{00}}^2$	by-item variance of I_{0i}	$\sim U(0, 3)$
${\omega_{11}}^2$	by-item variance of I_{1i}	$\sim U(0, 3)$
ρ_I	correlation between (I_{0i}, I_{1i}) pairs	$\sim U(-.8, .8)$
σ^2	residual error	$\sim U(0, 3)$
$p_{missing}$	proportion of missing observations	$\sim U(.00, .05)$

determined the population parameters $\beta_0, {\tau_{00}}^2, {\tau_{11}}^2, \rho_S, {\omega_{00}}^2, {\omega_{11}}^2, \rho_I$, and σ^2 by sampling from uniform distributions with ranges given in Table 2. We then simulated 24 subjects and 12 or 24 items from the corresponding populations, and simulated one observation for each subject/item pair. We also assumed missing data, with up to 5% of observations in a given data set counted as missing (at random).

For tests of Type I error, β_1 (the fixed effect of interest) was set to zero. For tests of power, β_1 was set to .8, which we found yielded power around 0.5 for the most powerful methods with close-to-nominal Type I error.

We considered nine different analyses, three variants of ANOVA and six variants of mixed-effects models (Table 3). All LMEMs were fit using the `lmer` function of the R package `lme4`, version 0.999375-39 (Bates et al., 2011), estimated using maximum likelihood estimation. Further information and `R` scripts can be found in an online appendix (http://talklab.psy.gla.ac.uk/simgen). When fitting LMEMs, an attempt was made to fit the specified model. However, in some cases, the estimation procedure did not converge. The assumption was that users of LMEMs would not simply give up in the face of nonconvergence, but instead explore simpler models. To this end, the random effects structure of the model was progressively simplified until convergence was attained. For between-items designs, this meant dropping the by-subjects random slope. For within-items designs, statistics from the partially converged model were inspected, and the slope associated with smaller variance was dropped. In the rare (.002%) of cases that

Table 3: Analyses performed on simulated datasets

Analysis	WSWI	WSBI	Test statistics
min-F'	X	X	min-F'
F_1	X	X	F_1
$F_1 \times F_2$	X	X	F_1, F_2
Random-intercepts LMEM	X	X	t, χ^2_{LR}, MCMC
Maximal LMEM	X	X	t, χ^2_{LR}
Forward-stepping LMEM, subjects then items	X	X	t, χ^2_{LR}
Forward-stepping LMEM, items then subjects	X		t, χ^2_{LR}
Backward-stepping LMEM, subjects then items	X		t, χ^2_{LR}
Backward-stepping LMEM, items then subjects	X		t, χ^2_{LR}

the random-intercept only model did not converge, the analysis was discarded.

There are various ways of obtaining p-values from LMEMs, and to our knowledge, there is little agreement on which method to use. Therefore, we considered three methods currently in practice: (1) treating the t-statistic as if it were a z statistic (i.e., using the standard normal distribution as a reference); (2) performing likelihood ratio tests, in which the deviance ($-2LL$) of a model containing the fixed effect is compared to another model without it but that is otherwise identical in random effects structure; and (3) using Markov Chain Monte Carlo (MCMC) sampling from the posterior distribution. Although (3) is the approach recommended by (Baayen et al., 2008), MCMC sampling is not implemented in `lme4` for models containing random correlation parameters. We therefore used (3) only for random intercept only models.

In addition to LMEMs with maximal random effects and random intercepts only, we examined the performance of stepwise-selection LMEMs, where the random effects structure was determined either by stepping forward from simple to more complex random effects or by stepping backward, from complex to simple random effects. The stepping process would terminate when comparison between the current model and the next model suggested the step was not warranted. These stepwise selection approaches used likelihood ratio tests to compare models, with the α level set to .05 for each comparison, which we assume is the standard approach in the literature. We tested both varieties where subject-effects steps were considered before item-effects steps and varieties where item-effects steps were considered before subject-effects. For between-item designs, there was only one possible step, and therefore forward and backward models are formally identical.

The within-item design, in contrast, has richer possibilities. For forward models, the stepwise procedure began with a random intercept model and terminated when a likelihood ratio test failed to indicate that including a given slope would improve the model, retaining the simpler model. If the first slope of the possible slopes did not "pass" the test, then the second slope was never tested. Conversely, the backward model began with the maximal random effects model and terminated when the likelihood ratio test showed that dropping the slope led to a significantly worse fit of the model, and the more complex model was retained. If the first slope to be tested was retained, the second slope was never tested. Any models that did not converge during forward or backward stepping were simply ignored, and the comparison would then be between the current model and the next model that converged.

We generated 100000 datasets for testing for each of the eight combinations (effect present/absent, between-/within-item manipulation, 12/24 items). The functions we used in running the simulations and processing the results are available in the `R` package `simgen`, which we have made available in a supplementary appendix, along with a number of `R` scripts using the package. The appendix also contains further information about the additional `R` packages and functions used for simulating the data and running the analyses.

Results and Discussion

An ideal statistical analysis method maximizes statistical power while keeping Type I error nominal (at the stated α level). Overall Type I error rates for the analyses are given in Table 4 for the between-item design and in Table 5 for the within-item design. The analyses in each table are (approximately) ranked, with analyses lower in the table showing the highest Type I error rates. Only min-F' was consistently at or below the stated α level. This is not entirely surprising, because the techniques that are available for deriving p-values from LMEMs with random slopes are known to be somewhat anticonservative (Baayen et al., 2008). For models with maximal random-effects structure or backward selection, this anticonservativity was quite minor, within 1–2% of α.[5] It is also worth noting

[5]This anticonservativity stems from underestimation of the variation between subjects and/or items, as is suggested by generally better performance of the maximal model in the 24- as opposed to 12-item simulations. In the appendix, we show that for LMEMs with random slopes Type I error decreases rapidly as additional subjects and items are added, while for RI-only models, error rate actually *increases*.

that $F_1 \times F_2$, which is known to be fundamentally biased (Clark, 1973; Forster & Dickinson, 1976), controlled overall Type I error better than forward-stepping or random-intercepts LMEMs, and almost as well as maximal LMEMs.

Table 4: Type I error rate for between-items design; RI-only = Random intercepts only.

	$\alpha = .01$		$\alpha = .05$		$\alpha = .10$	
N_{items}	12	24	12	24	12	24
	Type I Error at or near α					
min-F'	.009	.009	.044	.045	.092	.093
LMEM, Maximal, χ^2_{LR}	.017	.013	.070	.058	.129	.113
$F_1 \times F_2$	.014	.019	.063	.077	.120	.137
LMEM, Selection, χ^2_{LR}	.018	.013	.071	.058	.130	.113
	Type I Error far exceeding α					
LMEM, Maximal, t	.029	.017	.086	.065	.143	.120
LMEM, Selection, t	.030	.017	.088	.065	.145	.120
LMEM, RI-only, χ^2_{LR}	.032	.039	.102	.111	.171	.177
LMEM, RI-only, t	.055	.051	.128	.124	.193	.189
LMEM, RI-only, MCMC	.071	.103	.173	.211	.255	.294
F_1	.297	.217	.421	.339	.497	.420

F_1 alone was the worst performing test for between-items designs, and also had an unacceptably high error rate for within-items designs. LMEMs with random-intercepts only were also unacceptably anticonservative for both types of designs, far worse than $F_1 \times F_2$. In fact, for within-items designs, *random-intercepts-only LMEMs were even worse than F_1 alone*, showing false rejections 40-50% of the time at the .05 level, regardless of whether p-values were derived using the normal approximation to the t-statistic, the likelihood-ratio test, or MCMC sampling. Therefore, random-intercepts LMEMs represent a giant step backward in terms of simultaneous generalization over subjects and items.

Occupying an intermediate range were stepwise LMEMs. For between-items designs, where there was only one slope to make a decision about (the by-subjects random slope), using a stepwise procedure was not much different from using maximal random effects structure. However, performance of stepwise models degraded once there were two slopes to decide on, moreso for forward-stepping than for backward-stepping models. The worst performance was for forward stepping models. For 12-item datasets, forward models testing the subject slope first were most likely to end up at a random-intercepts model (about 23 % of cases) followed

Table 5: Type I error rate for within-items design; RI-only = Random intercepts only.

	α = .01		α = .05		α = .10	
N_{items}	12	24	12	24	12	24
	Type I Error at or near α					
min-F'	.004	.005	.027	.031	.061	.068
LMEM, Maximal, χ^2_{LR}	.013	.012	.059	.056	.113	.108
$F_1 \times F_2$	.012	.018	.057	.072	.112	.130
LMEM, Backward, Subjects First, χ^2_{LR}	.018	.012	.065	.058	.120	.110
LMEM, Backward, Items First, χ^2_{LR}	.019	.013	.067	.057	.123	.110
LMEM, Maximal, t	.022	.016	.072	.063	.126	.115
LMEM, Backward, Subjects First, t	.026	.017	.078	.064	.133	.117
LMEM, Backward, Items First, t	.027	.017	.080	.064	.135	.117
	Type I Error exceeding α					
LMEM, Forward, Items First, χ^2_{LR}	.056	.041	.117	.095	.176	.150
LMEM, Forward, Items First, t	.063	.045	.128	.101	.187	.156
LMEM, Forward, Subjects First, χ^2_{LR}	.076	.042	.140	.095	.200	.149
LMEM, Forward, Subjects First, t	.082	.046	.149	.101	.209	.155
F_1	.083	.059	.176	.139	.251	.210
LMEM, RI-only, χ^2_{LR}	.317	.377	.440	.498	.514	.567
LMEM, RI-only, t	.320	.379	.441	.499	.515	.568
LMEM, RI-only-MCMC	.260	.360	.390	.500	.440	.600

by forward models testing the item slope first (22 %) , and then backward models testing the subject or item slope first (4 % in each case).

From the point of view of overall Type I error rate, we can rank the analyses for both within- and between-items designs in order of desirability:

1. min-F' and maximal LMEMs;
2. backward-stepping LMEMs and $F_1 \times F_2$;
3. forward-stepping LMEMs;
4. F_1 and random-intercepts LMEMs.

It would also seem natural to draw a line separating analyses that have an "acceptable" rate of false rejections (i.e., 1–2) from those with a rate that is intolerably high (i.e., 3–4). However, it is insufficient to consider only the overall Type I error rate, as there may be particular problem areas of the parameter space where even the best analyses perform poorly (such as when particular variance components are very small or large). If these areas are small, they will only moderately affect the overall error rate. This is a problem because we do not know where the actual populations that we study reside in this parameter space; it could be that they inhabit these problem areas. It is therefore also useful to look at Type I error rate as a function of certain random effect parameters. This provides not only a further

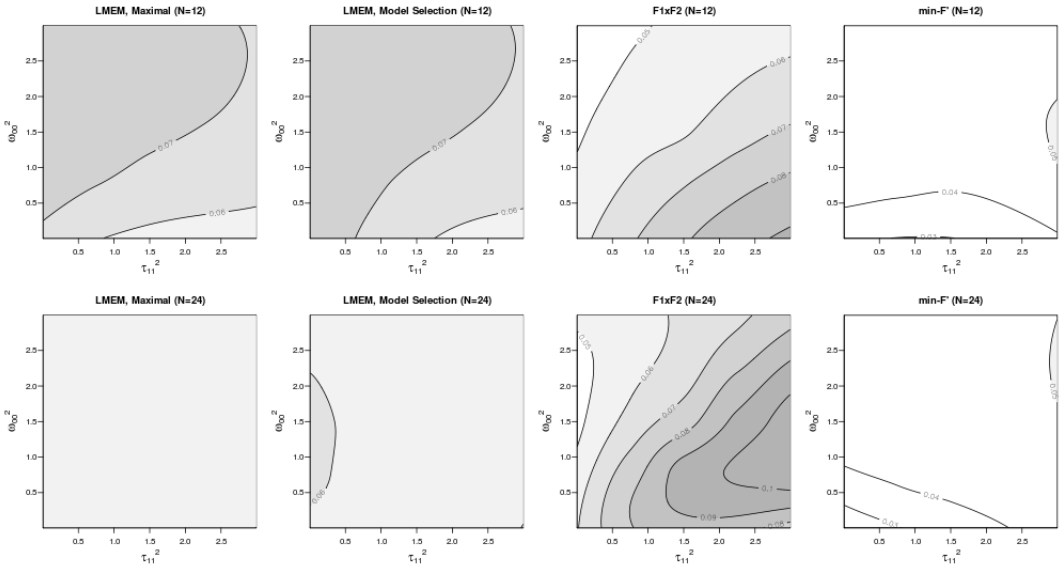

Figure 2: Type I Error rate for between-items design, analyses at or near the nominal α level, as a function of by-subject random slope variance ${\tau_{11}}^2$ and by-item random intercept variance ${\omega_{00}}^2$. The p-values for all LMEMs in the figure are from likelihood-ratio tests. Top row: 12 items; bottom row: 24 items.

check on the soundness of the best analyses, but also gives some insight into why some of the poorer analyses break down.

To probe deeper into these analyses' performances, it is useful to recognize the principles at play in determining statistical power for multilevel analyses (including mixed-model ANOVA). As with any inferential analysis, the theoretical limit on power ultimately derives from the signal-to-noise ratio. For multilevel models in particular, variability at the level of cluster (subject and/or item) is an essential part of this noise. Analyses that are fundamentally *sound* will interpret increased cluster-level variability as noise and thus will be more conservative. In contrast, analyses that are fundamentally *flawed* will tend to interpret this variability as signal, thus increasing the Type I error rate. The variances that are most in danger of being "misinterpreted" as signal are those that can drive differences between treatment means. Thus, in between-item designs, item variance is a potential culprit; in within-items designs, the item variance contributes equally to the two treatment means and is therefore factored out, but the treatment-by-item variance can drive differences between means. This implies that the variance components to look at for the within-subjects/between-items design are the by-item intercept variance

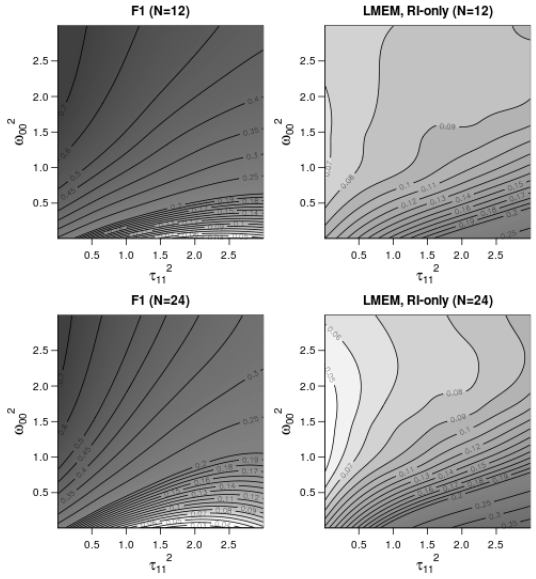

Figure 3: Type I Error rate for between-items design, analyses exceeding the nominal α level, as a function of by-subject random slope variance ${\tau_{11}}^2$ and by-item random intercept variance ${\omega_{00}}^2$. The p-values for all LMEMs in the figure are from likelihood-ratio tests. Top row: 12 items; bottom row: 24 items.

(${\omega_{00}}^2$) and the by-subject slope variance (${\tau_{11}}^2$); for the within-subjects/within-items design, they are the by-subject slope variance ${\tau_{11}}^2$ and the by-item slope variance ${\omega_{11}}^2$.

Figures 2–5 show the predicted Type I error rate (see supplementary on-line material for information) for the analyses as a function of these critical variances. From Figures 2 and 4 it can be seen that only min-F' maintains the Type I error rate consistently below the α-level throughout the parameter space. It can also be seen that min-F' becomes increasingly conservative as the relevant random effects become small, replicating Forster & Dickinson (1976). Maximal LMEMs show no such increasing conservativity, performing well overall, especially with 24-item datasets. In contrast, the performance of model selection approaches degrades as the critical random slope parameters become small, especially with respect to the slope that is tested as the first step in within-item designs (Figures 2 and 4). Random-intercepts LMEMs degrade extremely rapidly as a function of random slope parameters; even at very low levels of random-slope variability, the Type I error rate is unacceptably high (Figures 3 and 5).

The widely adopted $F_1 \times F_2$ criterion seems to occupy an interesting mid-

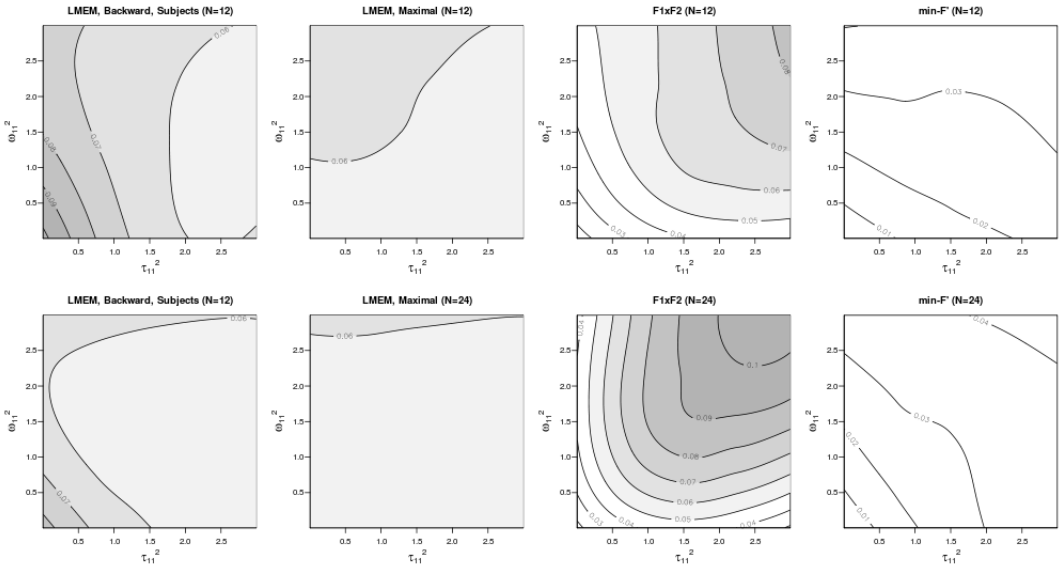

Figure 4: Type I Error rate for within-items design, analyses at or near the nominal α level, as a function of by-subject random slope variance ${\tau_{11}}^2$ and by-item random slope variance ${\omega_{11}}^2$. The p-values for all LMEMs in the figure are from likelihood-ratio tests. Top row: 12 items; bottom row: 24 items.

dle ground between maximal LMEMs and random-intercepts LMEMs. On the one hand, it is clearly far less anti-conservative than random-intercepts LME; in fact, its average behavior in terms of Type I error across our simulations is comparable to maximal LME analyses (slightly less anti-conservative for 12 items, slightly more anti-conservative for 24 items). The visualization in terms of the critical variances, however, reveals that $F_1 \times F_2$ is still fundamentally unsound: it becomes increasingly anti-conservative as either subject or item random slopes grow (Figure 4). As Clark (1973) pointed out, subject random slopes are not accounted for in the F_2 analysis, nor are item random slopes accounted for in the F_1 analysis. The fact that both F_1 and F_2 analyses have to pass muster keeps this anti-conservativity relatively minimal as long as subject and/or item slope variances are small, but the anti-conservativity is there nonetheless.

Note also that $F_1 \times F_2$ and backwards-stepping LMEMs yield almost complementary Type-I error distributions (Figure 4): while $F_1 \times F_2$ encounters problems when random slope variances get large, backward-stepping LMEMs encounter them when random slope variances get small. The latter is likely because backward selection more often eliminates the relevant slope terms from the model

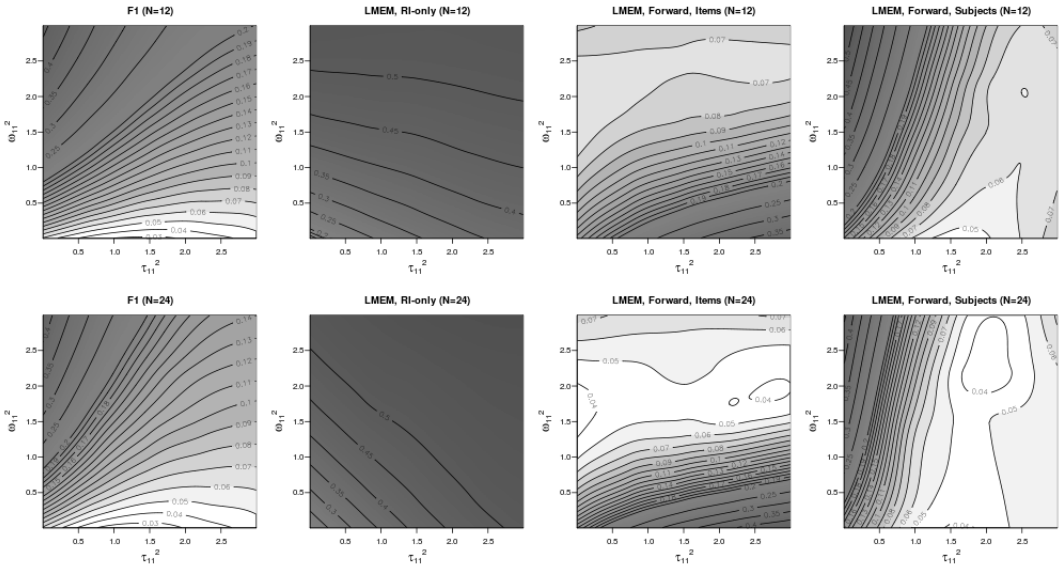

Figure 5: Type I Error rate for within-items design, analyses exceeding the nominal α level, as a function of by-subject random slope variance ${\tau_{11}}^2$ and by-item random slope variance ${\omega_{11}}^2$. The p-values for all LMEMs in the figure are from likelihood-ratio tests. Top row: 12 items; bottom row: 24 items.

(Type-II error due to lack of evidence for the inclusion of random slopes). Again, this underscores the general superiority of maximal LMEMs.

In sum, insofar as one is concerned about drawing conclusions likely to generalize across subjects and items, only min-F' and maximal LMEMs are fundamentally sound.[6] F_1-only and random-intercepts LMEMs are fundamentally flawed, as are forward selection models, especially in cases with few observations. The widely-used $F_1 \times F_2$ approach is flawed as well, but may be acceptable in cases where maximal LMEMs are not applicable. The question now is which of these analyses best maximizes power (Tables 6 and 7; Figures 6 and 7).

Overall, maximal LMEMs showed better greater power than min-F'; for the α=.05 level, it yielded power that was higher than min-F'. However, it is not immediately clear how much of this advantage is due to the greater anticonservativity of maximal LMEMs. One way to address this is to use the test statistics from the datasets we generated under the null hypothesis as a null-hypothesis dis-

[6]Backwards selection may also be categorizable as fundamentally sound, but we see nothing to recommend it over maximal LMEMs.

Table 6: Power for between-items design; RI-only = Random intercepts only.

	α = .01		α = .05		α = .10	
N_{items}	12	24	12	24	12	24
	Type I Error at or near α					
min-F'	.079	.154	.210	.328	.311	.444
LMEM, Maximal, χ^2_{LR}	.118	.185	.267	.364	.371	.478
$F_1 \times F_2$	.106	.222	.252	.403	.355	.510
LMEM, Model Selection, χ^2_{LR}	.119	.186	.269	.364	.372	.478
	Type I Error exceeding α					
LMEM, Maximal, t	.162	.214	.300	.382	.394	.490
LMEM, Model Selection, t	.164	.215	.302	.383	.395	.490
LMEM, RI-only, χ^2_{LR}	.164	.279	.319	.449	.419	.548
LMEM, RI-only, t	.228	.318	.360	.472	.447	.563
LMEM, RI-only-MCMC	.252	.444	.428	.601	.524	.680
F_1	.541	.571	.671	.706	.732	.767

tribution for the test statistics in the power analysis. This can be used to correct for the small degree of anticonservativity of maximal LMEMs. Applying these corrections, for the between-item design with α=.05, the corrected power values for maximal LMEMs (.223 and .342 for 12 and 24 items, respectively) were between 4% and 6% higher than the uncorrected values for min-F' (.210 and .328). Thus, there does not seem to be a substantial power advantage to using maximal LMEMs for between-item designs. In contrast, maximal LMEMs retained a considerable power advantage for within-items designs, with corrected power levels for α=.05 (.433 and .592) that were 16% to 32% higher than the uncorrected power values for min-F' (.327 and .512). Applying the same procedure to likelihood-ratio-test random-intercepts LMEMs reveals corrected powers *below* those of maximal LMEMs (between-items: .216 and .314 for 12 and 24 items respectively; within-items: .380 and .531). That is, most of the apparent additional power of maximal LMEMs over min-F' is real; but most of the apparent power of random-intercepts LMEMs is illusory.

General Discussion

Recent years have witnessed a surge in popularity of LMEMs in psycholinguistics and related fields. In many respects, the excitement these models have generated over the past several years is well deserved, given their great flexibility and their ability to model effects at the level of the individual trial. Despite this popularity and widespread use, there seems to be little understanding of the critical role of random effects in these models, leading to wildly varying random effects

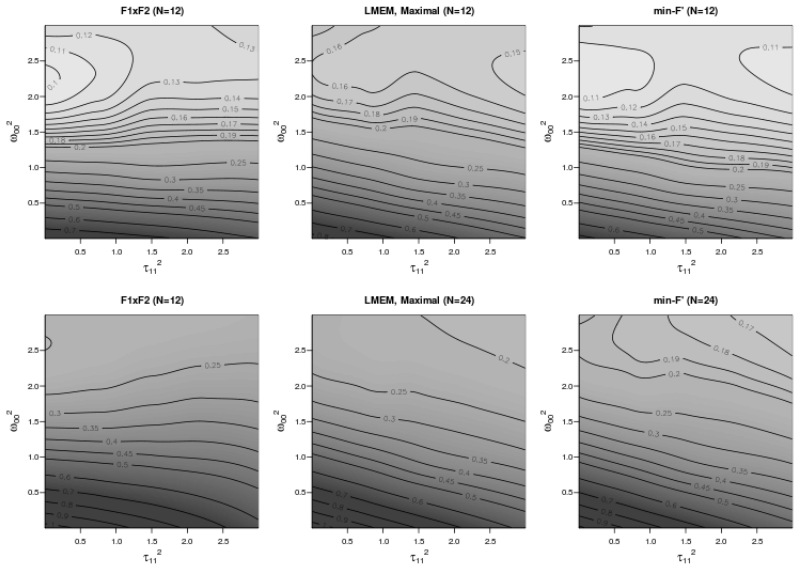

Figure 6: Power for between-items design, analyses at or near the nominal α level, as a function of by-subject random slope ${\tau_{11}}^2$ and by-item random slope ${\omega_{00}}^2$. The p-values for all LMEMs in the figure are from likelihood-ratio tests. Top row: 12 items; bottom row: 24 items.

specifications. We have emphasized that specifying random effects in an LMEM involves essentially the same principles involved in the selection of an analysis technique from the traditional menu of options. And since this is the case, researchers using LMEMs should adhere to the standards regarding generalization that have governed research in psycholinguistics and related fields for almost half a century. Historically, the standard for ANOVA has been to assume the presence of random condition-specific effects whenever observations across multiple conditions belong to a single cluster (most typically experimental subject and item, as explored in this paper). The reason for this is that our field is primarily concerned with discovering phenomena that generalize beyond the specific subjects or items involved in an experiment; if different subjects or items have different idiosyncratic sensitivities to experimental condition, then failing to include random condition-specific effects in ANOVA analysis often leads to inaccurate inferences about the generalizability of an observed effect. As we have shown both theoretically and through extensive simulations, this issue is every bit as true for LMEMs as for ANOVA.

Our survey of the generalizability of various analyses for data with crossed random effects leads to the clear conclusion that the only real contenders are max-

Table 7: Power for within-items design; RI-only = Random intercepts only.

	α = .01		α = .05		α = .10	
N_{items}	12	24	12	24	12	24
	Type I Error at or near α					
min-F'	.129	.266	.327	.512	.463	.643
LMEM, Maximal, χ^2_{LR}	.240	.382	.460	.610	.582	.717
$F_1 \times F_2$	.212	.410	.440	.643	.568	.746
LMEM, Backward, Subjects First, χ^2_{LR}	.251	.384	.467	.612	.586	.718
LMEM, Backward, Items First, χ^2_{LR}	.257	.384	.470	.612	.588	.718
LMEM, Maximal, t	.306	.425	.496	.629	.603	.727
LMEM, Backward, Subjects First, t	.315	.427	.502	.631	.608	.728
LMEM, Backward, Items First, t	.319	.427	.505	.631	.609	.728
	Type I Error exceeding α					
LMEM, Forward, Items First, χ^2_{LR}	.331	.429	.516	.635	.621	.733
LMEM, Forward, Items First, t	.384	.465	.547	.652	.640	.742
LMEM, Forward, Subjects First, χ^2_{LR}	.367	.428	.551	.635	.649	.734
LMEM, Forward, Subjects First, t	.412	.464	.575	.652	.664	.742
F_1	.455	.538	.640	.724	.725	.800
LMEM, RI-only, χ^2_{LR}	.787	.904	.853	.935	.883	.949
LMEM, RI-only, t	.789	.904	.854	.935	.883	.949
LMEM, RI-only-MCMC	.860	.898	.880	.918	.920	.939

imal LMEMs and min-F'. Given the lore about the conservatism of min-F' and the power of LMEMs, it comes as something of a surprise that, in terms of power, maximal LMEMs in between-items designs perform nearly identically to min-F', and in within-items designs shows a considerable improvement in power of between 16% and 32%.

Although min-F was the only analysis to consistently maintain the Type I error rate at or below α, there are be serious drawbacks to readopting it as a standard for the field. First, LMEMs offer numerous advantages over min-F', including being able to accommodate continuous predictors, and correct handling of imbalanced (e.g., missing) data. Another often-overlooked advantage is that it can accommodate richer clustering structure than just subjects and items as sampled units, which can be useful in studies of dialogue for example, where the dyad is often an important unit of analysis over and above the individual subject. Additionally, it is not yet clear whether min-F' performs well on other types of data, such as categorical data. Indeed, because min-F' is designed for continuous data, we suspect it would be vulnerable to the scaling artifacts that afflict conventional ANOVA when applied to categorical data from factorial designs (Jaeger, 2008). In contrast, maximal LMEMs can accommodate many different kinds of data without such artifacts. Finally, the inflation of the Type I error rate for maximal LMEMs with likelihood-ratio tests was minor (6-7% instead of 5%), and we feel this is

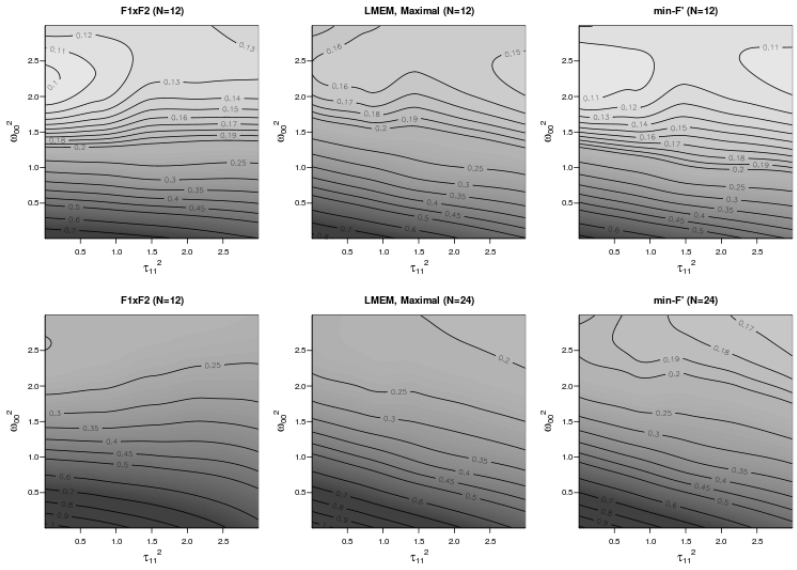

Figure 7: Power for within-items design, analyses at or near the nominal α level, as a function of by-subject random slope variance ${\tau_{11}}^2$ and by-item random slope variance ${\omega_{11}}^2$. The p-values for all LMEMs in the figure are from likelihood-ratio tests. Top row: 12 items; bottom row: 24 items.

a small price to pay for the benefit of having an approach that can more flexibly accommodate predictors and response variables of different types.

It is also clear from our survey that researchers who wish to use LMEMs need to be more attentive to how they specify random effects. Through this article, we have argued for a design-driven approach to specifying random effects. This contrasts, for example, with the common use of stepwise approaches in the literature. Although a stepwise approach might be justifiable in some circumstances—for example, when analyzing a very large corpus with many different possible models and many observations to boost power—it is certainly not justifiable in the case of a designed psychological experiment, where the possible sources of dependency can be clearly identified in advance. If the field wishes to loosen the standards for hypothesis testing, then it should do so across the board, not just for LMEMs. Indeed, it is possible to use a mixed-model ANOVA to decide whether or not a treatment-by-subject interaction or treatment-by-item interaction should be included in the model, and then to use simpler models if warranted. The fact that this approach has not emerged as a standard in the field might be taken to indicate that researchers have generally preferred analyses with maximal (or close to maximal) random effect structures.

Insofar as we as a field continue to be primarily concerned with discovering effects that generalize beyond specific experimental subjects and items, we should insist that researchers use random slopes in mixed-effects model analyses wherever they are justified by the design. Under this standard, random-intercepts LMEMs are only justifiable in two primary cases: (1) when all factors are-subject and between-items, and (2) in a within-subject/between-items design in which there is only a single item in each cell of the design.[7] In all other cases, *random-intercepts LMEMs are about the worst analysis one can perform*: they are definitely worse than $F_1 \times F_2$, and often even worse than F_1 alone!

Although in our investigation we have only looked at a very simple one-factor design with two levels and idealized data from a continuous response distribution, we see no reason why our results would not generalize to more complex designs and other types of data. First, the principles are the same in higher-order designs as they are for simple one-factor designs: any main effect or interaction for which there are multiple observations per subject or item can vary across these units, and, if this dependency is not taken into account, the p-values will be biased against the null hypothesis.[8] Furthermore, LMEMs generalize straightforwardly to other types of data such as categorical or count data through specification of a distribution function (e.g. binomial, poisson) and a link function (e.g., logit, probit). Because the underlying estimation algorithm is the same, we see little reason to expect the performance of LMEMs to differ much in these scenarios, although it is worth fully investigating this possibility. In contrast, we would expect the performance of ANOVA-based approaches to degrade because they assume that the response is normally distributed.

Producing generalizable results with LMEMs: Best practices

Our theoretical analyses and simulations lead us to the following set of recommended "best practices" for the use of LMEMs. We offer these not as the best

[7] Random-intercept models may also be appropriate in the analysis of visual-world eyetracking data and other time-series data when observations have been "aggregated up" to the subject or item level in order to minimize the impact of within-trial observational dependencies (Barr, 2008; Mirman et al., 2008). Still, the soundness of using random-intercept models in such cases should be further investigated, especially for higher order designs.

[8] To demonstrate this, we conducted Monte Carlo simulation of a 24-subject, 24-item 2×2 within/within experiment with main fixed effects, no fixed interaction, and random by-subject and by-item interactions. When analyzed with random-intercept LMEMs, we found a Type I error rate of .69; with maximal LMEMs the Type I error rate was .06. A complete report of these simulations appears in the supplementary appendix.

possible practices—as our understanding of these models is still evolving—but as the best given our current level of understanding.

Identifying the maximal random effects structure

As we have emphasized throughout this paper, the same considerations come into play when specifying random effects as when choosing from the menu of traditional analyses. So the first question you should ask yourself when trying to specify a maximal LMEM is: which factors are within-unit (subjects or items), and which are between? If a factor is between subjects or items, then the random intercept will do. Of course, the same principles apply for specifying by-item random slopes as for specifying by-subject random slopes, so to simplify the exposition we will only talk about the by-subject slopes. If a factor is within-subject, then chances are that you need a by-subject random slope for that effect. The only exception to this rule is when you only have a single observation for each combination of subject and treatment level. It may be the case that, due to missing data, some of your subjects (or items) have only one or zero observations for one or more treatment levels; still, one should at least try to estimate a random slope for this factor.

The same principles apply to higher-order designs involving interactions. In most cases, one should also have by-subject random slopes for any interactions where *all* factors comprising the interaction are within-subject; if any one factor involved in the interaction is between-subject, then the interaction cannot be estimated, and no random slope is needed. The exception to this rule, again, is when you have only one observation per subject per cell[9]. If some of the cells for some of your subjects have only one or zero observations, you should still try to fit a random slope.

Random effects for control predictors

One of the most compelling aspects of mixed-effects models is the ability to include almost any control predictor—by which we mean a property of an experimental trial which may affect the response variable but is not of theoretical interest in a given analysis—desired by the researcher. In principle, including control variables in an analysis can rule out potential confounds and increase statistical power by reducing residual noise. Given the investigations in the present paper, however, the question naturally arises: in order to guard against anti-conservative

[9] A cell is any unique combination of all the factors involved in the interaction; in a 2x2 design, where both factors are within subject, there are four cells.

inference about a predictor X of theoretical interest, do we need by-subject and by-item random effects for all our control predictors C as well? Suppose, after all, if there is no underlying fixed effect of C but there is a random effect of C—could this create anti-conservative inference in the same way as omitting a random effect of X in the analysis could? To put this issue in perspective via an example, Kuperman et al. (2010) include a total of eight main effects in an LME analysis of fixation durations in Dutch reading; for the interpretation of each main effect, the other seven may be thought of as serving as controls. Fitting eight random effects, plus correlation terms, would require estimating 72 random effects parameters, 36 by-subject and 36 by item. One would likely need a huge dataset to be able to estimate all the effects reliably (and one must also not be in any hurry to publish, for even with huge amounts of data such models can take extremely long to converge).

To our knowledge, there is little guidance on this issue in the existing literature, and more thorough research is needed. Based on a limited amount of informal simulation, however (reported in the supplementary appendix), we propose the working assumption that it is not essential for one to specify random effects for control predictors to avoid anticonservative inference, as long as interactions between the control predictors and the the factors of interest are not present in the model (or justified by the data).

Coping with failures to converge

It is altogether possible that the maximal LMEM will not converge with the full random-effects specification. In our experience, the likelihood that a model will converge depends on two factors: (1) the extent to which random effects in the model are large, and (2) the extent to which there are sufficient observations to estimate the random effects. Generally, as the sizes of the subject and item samples grow, the likelihood of convergence will increase. Of course, one does not always have the luxury of using many subjects and items. And, although the issue seems not to have been studied systematically, it is our impression that fitting maximal LMEMs is less often successful for categorical data than for continuous data.

It is fortunate that LMEMs will be more likely to converge when the random effects are large, since this is exactly the situation where $F_1 \times F_2$ is anti-conservative. This points toward a possible practice of trying to fit a maximal LMEM wherever possible, and resorting to $F_1 \times F_2$ analyses if the model will not converge and if ANOVA is appropriate for the design. It is important, however, to resist the temptation to step back to random-intercepts models. When the maximal

LMEM does not converge, the first step should be to check for possible misspecifications or data problems that might account for the error. A common mistake is to specify an unidentifiable model by including effects that cannot be estimated from the data such as a by-item random slope for a between-item effect. It may also help to use standard outlier removal methods and to center or sum-code the predictors. Once data and model specification problems have been eliminated, the next step is to seek out the next most complex model that does converge. We recommend the following approach. When using a package such as `lme4`, it is possible to inspect the random effects from the nonconverged model. One can then identify the highest-order term whose variance is the smallest, remove that term, and then re-fit the model. This process can be repeated until convergence is achieved.[10]

However, a cautionary note may be in order here. A situation may arise where the above strategy produces 'theoretically undesirable' models in which, for example, a higher order term is dropped that might be essential for distinguishing between two alternative hypotheses. We are not entirely sure how to proceed in such cases, and clearly, the issue of non-convergence needs to be addressed more fully in future research. A potential last resort in case of severe convergence problems might be to drop the concept of crossed random effects altogether and perform separate by-subject and by-item LMEMs, similar in logic to $F_1 \times F_2$, each with appropriate maximal random effect structures, or possibly after aggregating up to confound random slope variance with residual error, thus enabling the use of random-intercepts LMEM, although the soundness of this approach should be further investigated.

Computing p-values

There are a number of ways for computing p-values from LMEMs, none of which is uncontroversially the best. Although Baayen et al. (2008) recommended estimating them through Monte Carlo Markov Chain (MCMC) simulation, to our knowledge this is not yet implemented for maximal LMEMs in easily accessible software packages.[11] Our simulations suggest that until a more general MCMC

[10] We evaluated the viability of this recommendation by correlating the random slope ranks from the nonconverged within-subject/within-item LMEMs with the actual ranks from the relevant generative models. The correlations obtained were reasonably high at $r = .74$ ($N = 387$) and $r = .83$ ($N = 192$) for 12-item and 24-item experiments, respectively.

[11] MCMC simulations for random-slopes and more complex mixed-effects models can be run with general-purpose graphical models software such as WinBUGS (Lunn et al., 2000) or JAGS

solution arrives, we consider the next best approach for typical psycholinguistic datasets—where the number of observations far outnumbers the number of model parameters—the likelihood-ratio (LR) test. To perform such a test, one compares a model containing the fixed effect of interest to a model that is identical in all respects except the fixed effect in question. One should not also remove any random effects associated with the fixed effect when making the comparison. In other words, LR tests of a fixed effect with k levels should have only $k - 1$ degrees of freedom (e.g., one degree of freedom for the dichotomous single-factor studies in our simulations). We have seen cases where removing the fixed effect causes the comparison model to fail to converge. Under these circumstances, one can alter the comparison model following the procedures described above to attempt to get it to converge, and once convergence is achieved, compare it to an identical model including the fixed effect. Note that our results indicate that the concern voiced by Pinheiro & Bates (2000) regarding the anti-conservativity of likelihood-ratio tests to assess fixed effects in LMEMs is essentially unfounded, at least for datasets of the typical size of a psycholinguistic study.

Reporting results

It is not only important for researchers to understand the importance of using maximal LMEMs, but also for them to articulate their modeling efforts with sufficient detail so that other researchers can understand and replicate the analysis. In our informal survey of papers published in JML, we sometimes found nothing more than a mere statement that researchers used "a mixed effects model with random effects for subjects and items." This could be anything from a random-intercepts only to a maximal LMEM, and obviously, there is not enough information given to assess the generalizability of the results. One needs to provide sufficient information to the reader to be able to recreate the analyses. One way of satisfying this requirement is to report variance-covariance matrix, which includes all the information about the random effects, including their estimates. This is useful not only as a check on the random effects structure, but also for future meta-analyses, etc. A simpler option is to mention that one attempted to use a maximal LMEM and, as an added check, also state which factors had random slopes associated with them. If a slope was needed but excluded to attain convergence, this should also be stated, as well as the decision criteria for removing

(Plummer, 2003); in particular JAGS has a good interface to `R` called `rjags` which can be useful for this purpose. This approach can be delicate and error-prone, however, and we do not recommend it at this point as a general practice for the field.

slopes. For example: "We used a maximal LMEM with by-subject random slopes for factors X, Y, and by-item random slopes for factor X. Although required, the maximal LMEM including a by-subject slope for the X-by-Y interaction did not converge. Inspection of the random effects parameter estimates from the unconverged model suggested that the X-by-Y interaction term was the highest-order slope with the smallest variance, and so was dropped from the model."

At a recent workshop on mixed-effects models, a prominent psycholinguist[12] memorably compared encouraging psycholinguists to use linear mixed-effects models to giving shotguns to toddlers. Might the field be better off without complicated mixed-effects modeling, and the potential for misuse they bring? Although we acknowledge this complexity and its attendant problems, we feel that one of the reasons why researchers have been using mixed-effects models incorrectly is due to the misconception that they are something entirely new, a misconception that has prevented us from seeing the continued applicability of their previous knowledge about what a generalizable analysis requires. As we hope to have shown, by and large, *researchers already know most of what is needed* to use LMEMs appropriately. So long as we can continue to adhere to the standards that are already implicit, we therefore should not deny ourselves access to this new addition to the statistical arsenal. After all, when our investigations involve stalking a complex and elusive beast (whether the human mind or the feline palate), we need the most powerful weapons at our disposal.

Acknowledgements

We thank Ian Abramson and Simon Garrod for helpful feedback and suggestions. This work has been presented at seminars and workshops in Edinburgh, York, and San Diego in Spring 2011.

References

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

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

[12] G.T.M. Altmann

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Davenport, J. M., & Webster, J. T. (1973). A comparison of some approximate F-tests. *Technometrics*, *15*, 779–789.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Forster, K., & Dickinson, R. (1976). More on the language-as-fixed-effect fallacy: Monte carlo estimates of error rates for F1,F2,f, and min f. *Journal of Verbal Learning and Verbal Behavior*, *15*, 135–142.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, MA: Cambridge University Press.

Goldstein, H. (1995). *Multilevel Statistical Models* volume 3 of *Kendall's Library of Statistics*. Arnold.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, *59*, 434–446.

Jaeger, T. F. (2011). Post to the R-LANG mailing list, 20 February 2011, http://pidgin.ucsd.edu/pipermail/r-lang/2011-February/000225.html.

Kuperman, V., Bertram, R., & Baayen, R. H. (2010). Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language*, *62*, 83–97.

Locker, L., Hoffman, L., & Bovaird, J. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*, *39*, 723–730.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

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

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication*, *43*, 103–121.

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

Roland, D. (2009). Relative clauses remodeled: The problem with mixed-effect models. Poster presentation at the 2009 CUNY Sentence Processing Conference.

Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using Quasi *F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin*, *86*, 37–46.

Scheffe, H. (1959). *The analysis of variance*. New York: Wiley.

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Wickens, T. D., & Keppel, G. (1983). On the choice of design and of test statistic in the analysis of experiments with sampled materials. *Journal of Verbal Learning and Verbal Behavior*, *22*, 296–309.