

# Parsimonious Mixed Models

Douglas Bates

*Statistics, University of Wisconsin-Madison, Madison, USA.*

E-mail: bates@stat.wisc.edu

Reinhold Kliegl

*Psychology, University of Potsdam, Potsdam, Germany.*

Shravan Vasishth

*Linguistics, University of Potsdam, Potsdam, Germany,*

R. Harald Baayen

*Linguistics, University of Tübingen, Tübingen, Germany*

*Linguistics, University of Alberta, Edmonton, Canada*

**Summary.** The analysis of experimental data with mixed-effects models requires decisions about the specification of the appropriate random-effects structure. Recently, Barr, Levy, Scheepers, and Tily 2013 recommended fitting ‘maximal’ models with all possible random effect components included. Estimation of maximal models, however, may not converge. We show that failure to converge typically is not due to a suboptimal estimation algorithm, but is a consequence of attempting to fit a model that is too complex to be properly supported by the data, irrespective of whether estimation is based on maximum likelihood or on Bayesian hierarchical modeling with uninformative or weakly informative priors. Importantly, even under convergence, overparameterization may lead to uninterpretable models. We provide diagnostic tools for detecting overparameterization and guiding model simplification.

**Keywords:** linear mixed models, model selection, crossed random effects, model simplicity

## 1. Introduction

During the last ten years, there has been a significant change in how psycholinguistic experiments are analyzed when both subjects and items are included as random factors, specifically a change from analyses of variance to linear mixed models (LMMs), with Baayen et al. (2008) providing a first major introduction. Although hierarchical linear models have been in existence for decades, their adoption in areas like psychology and linguistics became more widespread in areas like linguistics and psychology after Pinheiro and Bates (2000) appeared (Vasishth, 2003; Vasishth and Lewis, 2006; Baayen et al., 2008; Oberauer and Kliegl, 2006; Kliegl et al., 2007; Kliegl, 2007). Recently, the use of LMMs has spread to other areas of psychology, such as personality and social psychology (Judd et al., 2012; Westfall et al., 2014). There are a number of reasons for this change. One particularly attractive feature has been that, with LMMs, statistical inference about experimental effects and interactions no longer needs separate analyses of variance, one

for subjects and one for items (Clark, 1973; Forster and Dickinson, 1976), but can be carried out within a single coherent framework.

This benefit in coherence comes at some cost. An important part of analyzing experimental data with mixed-effects models is the selection of the proper random-effects structure. In principle, LMMs not only consider variance between subjects and between items in the mean of the dependent variable (i.e., random intercepts), but also variance between subjects and between items for all main effects and interactions (i.e., random slopes) as well as correlations between intercepts and slopes. Let us illustrate the basic point with the most simple model with a two-level within-subject experimental manipulation and subject as the only random factor, using the formula notation of the ‘lme4’ package for R described in Bates et al. (2015a):

```
Y ~ 1 + A + (1|Subject)
```

This model may support the fixed effect of the within-subject factor A. However, if the effect of A (i.e., the difference between the two experimental conditions) differs reliably between subjects, uncertainty about A may be so substantial that the main effect of A is no longer significant in a model allowing for random slopes for A:

```
Y ~ 1 + A + (1+A|Subject)
```

For the assessment of the significance of the experimental manipulation, it is therefore essential to examine its fixed effect in the presence of the corresponding random slopes (see, e.g., Pinheiro and Bates (2000), Baayen (2008)).

## 2. Parameter estimation in mixed models

Following the publication of Barr et al. (2013) containing the advice in its title to “keep it maximal” when formulating an LMM for confirmatory analysis, the frequency of reports and queries related to failure of a model fit to converge has increased, for example, in the discussion list for the `lme4` package for R (Bates et al., 2015b). Although LMMs and generalized linear mixed-effects models (GLMMs) are versatile tools for modeling the variability in observed responses and attributing parts of this variability to different sources, like any statistical modeling technique they have their limitations. Knowledge of these limitations is important in ensuring appropriate usage.

As the complexity of the model increases, so does the difficulty of the optimization problem. For LMMs, aside from intercept and fixed effects, the parameters being estimated represent variances and covariances, which are typically much more difficult to estimate than regression coefficients. The parameters in GLMMs are even more complicated.

When there is more than one experimental factor, say in a factorial design, the number of parameters to be estimated explodes. For example, a maximal model with three experimental within-subject and within-items factors with two levels each in a full factorial design incorporating the seven main effects and interactions plus the intercept with random effects for subject and item, would require estimation of eight fixed-effects coefficients and 73 variance-covariance parameters. The online supplement to Barr et al.

(2013) fits such a model to data from an experiment described in Kronmüller and Barr (2007). We also present a reanalysis of this experiment below.

To anticipate the main result, it is simply not realistic to try to fit this number of highly abstract parameters given the number of subjects and items in this experiment. Almost unfortunately, the software does indeed converge to parameter estimates but these estimates correspond to degenerate or singular covariance matrices, in which some linear combinations of the random effects are estimated to having no variability. This corresponds to estimates of zero random-effects variance in a model with random-intercepts only or a correlation of  $\pm 1$  in a model with correlated random intercepts and slopes. However, already a three-by-three correlation matrix will not usually show boundary values like these, even when it is singular. In summary, the parameters representing variances and covariances are constrained in complicated ways. In overparameterized models, convergence can occur on the boundary, corresponding to models with singular variance-covariance matrices for random effects. This can have serious, adverse consequences for inference; for example, due to an overparameterization of the maximal LMM, Kliegl et al. (2011) wrongly interpreted an LMM correlation parameter as providing much more evidence than the corresponding within-subject correlation for the correlation of two experimental effects.

In a linear mixed model incorporating vector-valued random effects, say by-subject random effects for intercept and for slope, the variance component parameters determine a variance-covariance matrix for these random effects. As described in Bates et al. (2015a), the parameters used in fitting the model are the entries in the Cholesky factor ( $\Lambda$ ) of the relative variance-covariance matrix of the unconditional distribution of the random effects. The parameter vector ( $\theta$ ) for this model are the values on and below the diagonal of a lower triangular Cholesky factor. The  $\theta$  vector elements fill the lower triangular matrix in column major order. The relative covariance matrix for the random effects is  $\Lambda\Lambda'$ . To reproduce the covariance matrix,  $\Lambda\Lambda'$  must be scaled by  $s^2$ .

When one or more columns of the Cholesky factor  $\Lambda$  are zero vectors,  $\Lambda$  is rank-deficient: The linear subspace formed by all possible linear combinations of the columns is of reduced dimensionality compared to the dimensionality of  $\Lambda$ . The random-effects vectors that can be generated from this fitted model must lie in this lower-dimensional subspace. That is, there will be no variability in one or more directions of the space of random effects.

The **RePsychLing** package provides a new function, **rePCA** (which may become part of a future release of 'lme4') that enables the analyst to probe models fitted with **lmer** for rank deficiency. The **rePCA** (random-effects Principal Components Analysis) function takes an object of class **lmerMod** (i.e. a model fit by **lmer**) and produces a list of principal component (**prcomp**) objects, one for each grouping factor. These principal component objects can be summarized and visualized (by means of scree plots), exactly as any other principal component object generated by the **prcomp** function of R.

Principal components analysis of the estimated covariance matrices for the random effects in a linear mixed model allows for simple assessment of the dimensionality of the random effects distribution. As illustrated below and in the vignettes in the **RePsychLing** package, the maximal model in many analyses of data from Psychology and Linguistics experiments, is almost always shown by this analysis to be degenerate.

### 3. Iterative reduction of model complexity

In this section, we assume that the researcher has hypotheses about main effects and (some) interactions, but that he/she has no specific expectations about variance components or correlation parameters. In other words, we assume that the experimental hypotheses relate to the fixed effects, not the random-effects structure. In hypothesis testing, usually the primary reason for dealing with the random-effects structure is to obtain as powerful tests as justified of the fixed effects. Therefore, it is reasonable to remove variance components/correlation parameters from the model if they are not supported by the data. If there are specific expectations about, say, a correlation parameter, it makes sense to include it in the model (as well as the related variance components). We also assume the standard situation whereby potential numeric within-subject or within-item covariates are not under consideration.

In a factorial experiment, the maximum number of variance-covariance parameters to be estimated for each random factor is

$$\frac{(\text{product of within-factor levels}) \times (\text{product of within-factor levels} + 1)}{2}.$$

For example, a  $2 \times 2$  within-factor design will have  $(2 \times 2) \times ((2 \times 2) + 1)/2 = 10$  parameters in the variance-covariance matrix for each random effect (commonly, subject and item), and a  $2 \times 2 \times 2$  within-factor design will have 36 parameters. Additional model parameters are required for the fixed effects intercept, main effects, interactions, and for the residual variance. It is reasonable to start by attempting to fit a maximal LMM. If this model converges within reasonable time, several steps can be taken to check the possibility of an iterative reduction of model complexity in order to arrive at a parsimonious LMM.

First, it is worth checking whether we can reduce the dimensionality of the variance-covariance matrices assumed in a maximal LMM. The number of principal components that cumulatively account for 100% of the variance is a reasonably stringent criterion for settling on the reduced dimensionality. This can be achieved by performing PCA using the `rePCA()` function on the fitted maximal LMM (see section 2 above).

Second, after we have determined the number of dimensions supported by the data, we can eliminate variance components from the LMM, following the standard statistical principle with respect to interactions and main effects: variance components of higher-order interactions should generally be taken out of the model before lower-order terms nested under them. Frequently, in the end, this leads also to the elimination of variance components of main effects. The reduced model may be submitted again to a PCA to check the dimensionality of the random-effects structure.

Third, we can check whether forcing to zero the correlation parameters of the reduced LMM significantly decreases the goodness of fit according to a likelihood ratio test (LRT), possibly also taking into account changes in AIC and BIC. Obviously, if the goodness of fit does not change from the reduced model to the zero-correlation-parameter (ZCP) model, we do not have reliable evidence that the correlation parameters are different from zero. Importantly, this does not mean that the correlations are zero, only that we do not have enough evidence for them being different from zero for the current data; absence of evidence is not evidence of absence. Also, note that the estimated value

of the correlation parameters depends on the choice of contrasts for the experimental factors. For example, treatment contrasts and sum contrasts may lead to models with a very different random-effect structure (see <http://www.rpubs.com/Reinhold/22193>; this is also available as a vignette in the `RePsychLing` package). It is also conceivable that correlation parameters are zero for only one of several random factors. The new `'||'` syntax of `'lmer()'` is very convenient for specifying such ZCP LMMs. However, as illustrated in the same vignette, there are a few constraints relating to general R-formula syntax one needs to know about. In general, the `'||'` syntax works (currently) as expected only for LMMs after converting factor-based to vector-valued random-effects structures.

Fourth, we may also want to check whether all the variance components of an identified model are necessary. Taking out one term at a time and checking again whether there is a significant drop in goodness of fit, allows us to identify variance components that are not supported by the data. Again, removing such terms does not mean that the variance is zero, only that we have no evidence of it being significantly different from zero.

Fifth, after removing non-significant variance components, we may want to recheck whether the goodness of fit of the iteratively reduced model increases if it is extended with correlation parameters. A reliable variance parameter (i.e., a variance parameter contributing significantly to the goodness of fit according to a likelihood-ratio test) is a necessary condition for estimating correlation parameters associated with this variance component. In other words, we expect to find statistically significant correlation parameters only if the related variance components are statistically significant by themselves.

Finally, as a special case, the iterative reduction of model complexity described above assumed that there was a solution for the maximal model. With complex experiments, however, it may happen that the maximal model does not converge to a solution (e.g., there are warnings that no solution was found) or that the solution is obviously degenerate (i.e., variance components or correlation parameters are estimated at their boundaries of 0 or  $\pm 1$ , respectively). In this case, a first step could be to check the dimensionality of the zero-correlation parameter model with a PCA. Obviously, with complex experiments the switch from maximal to zero-correlation parameter model will yield the largest simplification. As already mentioned, whenever one switches from maximal to zero-correlation parameter models, it is very important to have a clear understanding of the contrast specifications chosen for the experimental factors.

In the following two sections, we describe iterative reductions of LMM complexity for two experiments using the above checks. The data for the first example are from an experiment on pragmatic comprehension of instructions (Kronmüller and Barr, 2007, Exp. 2; reanalyzed with an LMM in Barr et al., 2013). The second data set is from a visual-attention experiment (Kliegl et al., 2015), following up a different report with the overparameterized model (Kliegl et al., 2011). Detailed reports (including R code) for these analyses are described in vignettes, along with four additional examples. We emphasize that we do not claim that our illustrations are the only way to carry out these analyses, but the strategy outlined above has yielded satisfactory results with all data sets we have analyzed so far. There is no cook-book substitute for theoretical considerations and developing statistical understanding. Each data-set deserves the exercise of judgement on part of the researcher.

### 3.1. *Reanalysis of Kronmüller and Barr (2007)*

Here we apply the iterative reduction of LMM complexity to truncated response times of a  $2 \times 2 \times 2$  factorial psycholinguistic experiment (Kronmüller and Barr, 2007). This is their Exp. 2, reanalyzed with an LMM in Barr et al. (2013). The data are from 56 subjects who responded to 32 items. Specifically, subjects had to select one of several objects presented on a monitor with a cursor. The manipulations involved (1) auditory instructions that maintained or broke a precedent of reference for the objects established over prior trials, (2) with the instruction being presented by the speaker who established the precedent (i.e., an old speaker) or a new speaker, and (3) whether the task had to be performed without or with a cognitive load consisting of six random digits. All factors were varied within subjects and within items. There were main effects of Load (L), Speaker (S), and Precedent (P); none of the interactions were significant. Although standard errors of fixed-effect coefficients varied slightly across models, our re-analyses afforded the same statistical inference about the experimental manipulations as the original article, irrespective of LMM specification (see Figure 1 a comparison of fixed effects of maximal and parsimonious LMMs). The purpose of the analysis is to illustrate an assessment of model complexity as far as variance components and correlation parameters are concerned, neither of which were the focus of the original publication.

#### 3.1.1. *Maximal linear mixed model*

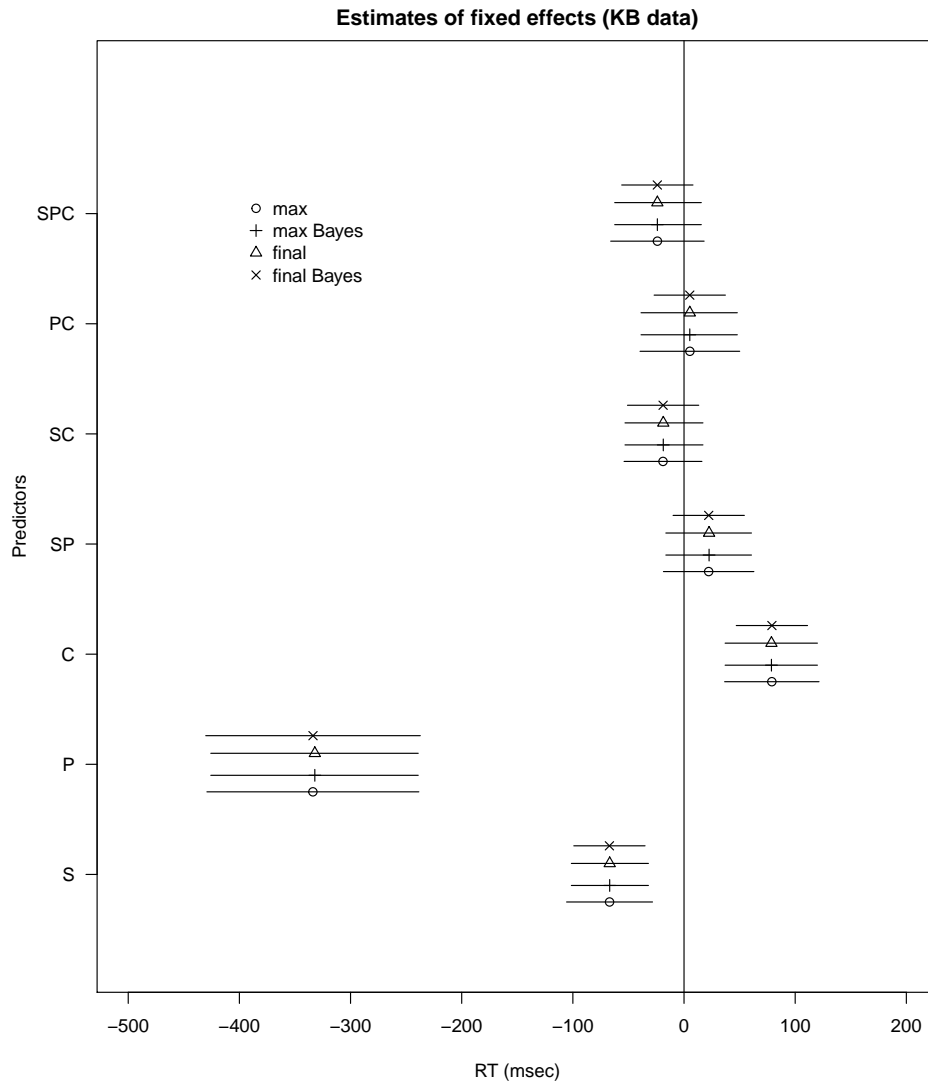
A full factorial model in the fixed-effects can be described by the formula

```
~1+S+P+C+SP+SC+PC+SPC
```

Barr et al. (2013) analyzed Kronmüller and Barr (2007, Exp. 2) with the maximal model for this design comprising 16 variance components (eight each for the random factors `SubjID` and `ItemID`, respectively). The model took 39,004 iterations to converge, but produces what look like reasonable parameter estimates (i.e., no variance components with estimates close to zero; no correlation parameters with values close to  $\pm 1$ ). The slow convergence is due to the total of  $2 \times 36 = 72$  parameters in the optimization of the random-effects part (ignoring the eight fixed-effect parameters and the residual variance, which do not contribute much to the computational load). Figures 1 and 2 display the fixed effects and variance components of the maximal model, along with other estimates, discussed below. The correlations of subject and item random effects are shown in Figures 3 and 4.

Considering that there are only 56 subjects and 32 items, it is quite optimistic to expect to estimate 36 covariance parameters for `SubjID` and another 36 for `ItemID`. A principal components analysis of the variance-covariance matrices for subject and item random effects returns eight principal components, along with the cumulative proportions of variance explained (see Table 1). For subject random effects, four dimensions are sufficient to account for 100% of the variance explained; and for items, five dimensions suffice.

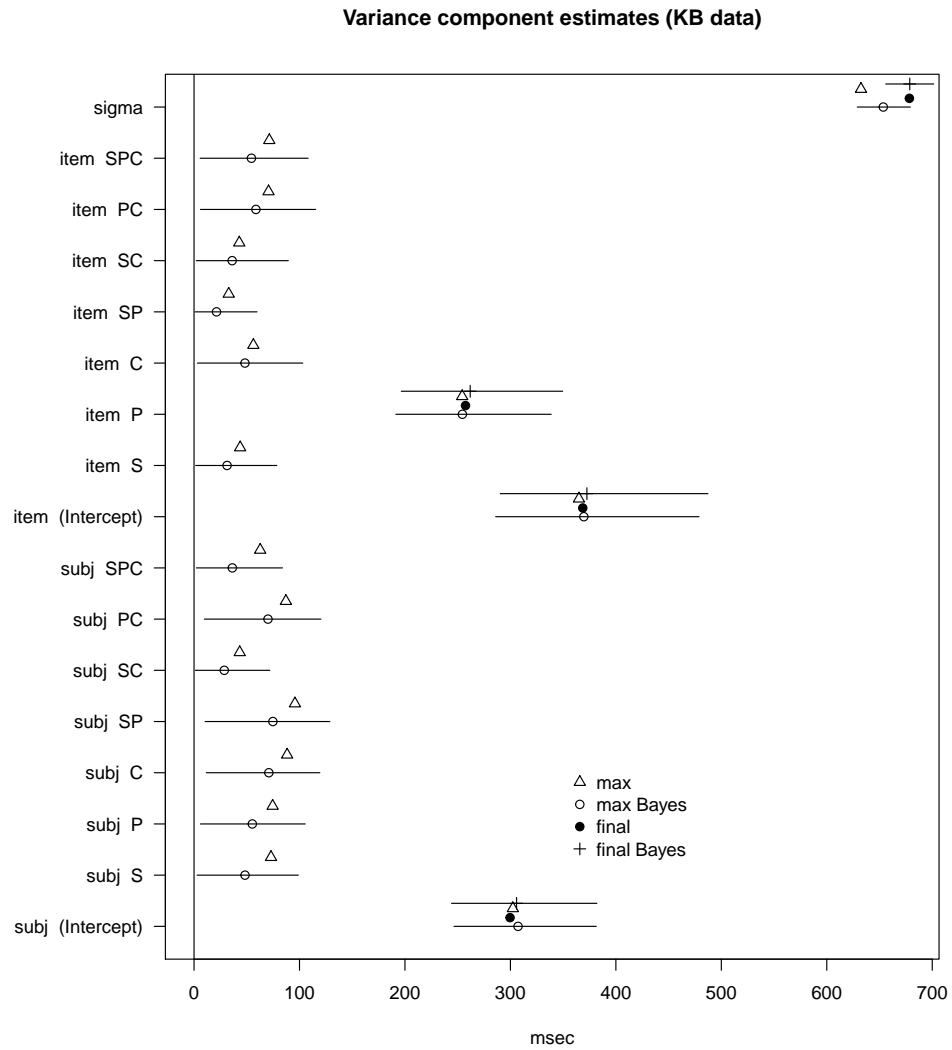
Thus, the maximal model is clearly too complex. In the following paragraphs, we illustrate our iterative method that reduces model complexity to arrive at an optimal LMM for this experiment. We will not report the intermediate results here, but they



**Fig. 1.** The estimates and 95% confidence intervals for the fixed effects in the maximal and final models of the Kronmüller and Barr 2007 data. Also shown are estimates and 95% credible intervals from a maximal Bayesian hierarchical linear model.

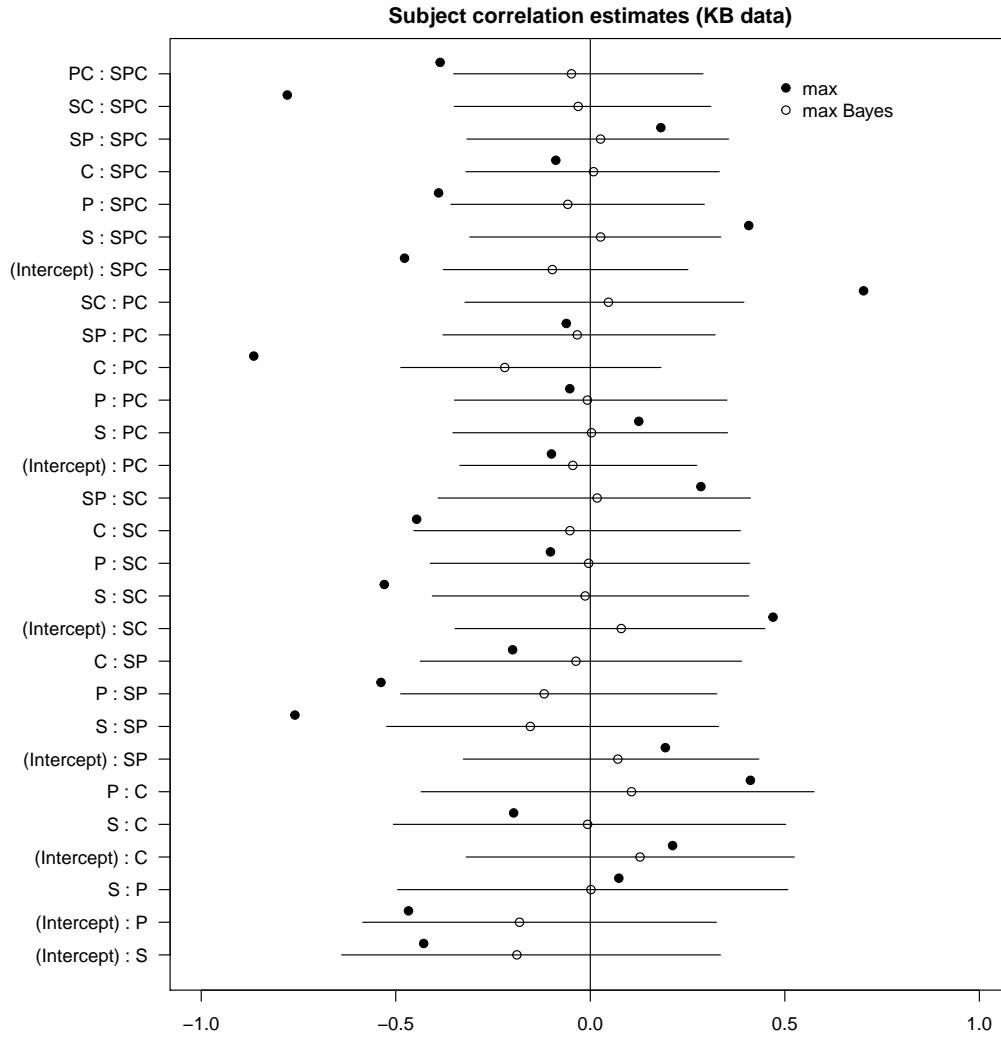
**Table 1.** The cumulative proportion of variance explained for the subject and item random effects in the maximal model for the Kronmüller and Barr 2007 data. Principal components analysis was used to compute the cumulative proportion of variance explained.

		1	2	3	4	5	6	7	8
subject	cum. prop.	0.73	0.85	0.94	1.00	1.00	1.00	1.00	1.00
item	cum. prop.	0.79	0.94	0.97	0.99	1.00	1.00	1.00	1.00

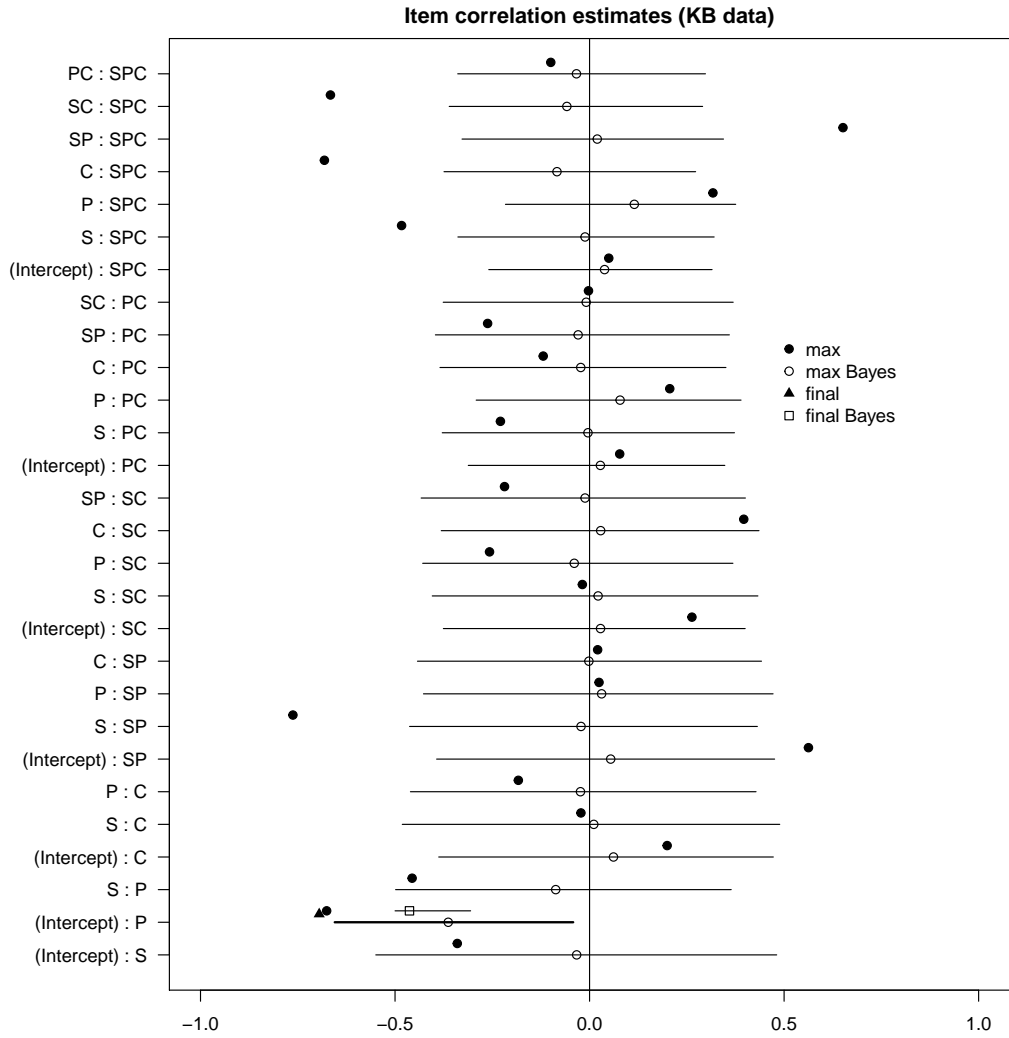


**Fig. 2.** The estimates of the variance components (standard deviations) from the maximal and final models of the KB data. Also shown are the estimates and 95% credible intervals of the maximal Bayesian hierarchical model.





**Fig. 3.** The estimates of the subject random effect correlations in the lme4 and Bayesian maximal models of the Kronmüller and Barr 2007 data. The Bayesian estimates also have 95% credible intervals.



**Fig. 4.** The estimates of the item random effect correlations in the lme4 and Bayesian maximal models of the Kronmüller and Barr 2007 data. The Bayesian estimates also have 95% credible intervals.

are available in the vignettes, along with the R code. We qualify this procedure at the outset: We do not claim that this is the only way to proceed, but the strategy has consistently yielded satisfactory results for all data sets we have examined so far.

### 3.1.2. Zero-correlation-parameter linear mixed model

As a first step toward model reduction, we start with a model including all 16 variance components, but no correlation parameters. The PCA of this model shows that 12 out of 16 dimensions suffice for capturing 100% of the variance explained. This suggests that the model is still too complex.

### 3.1.3. Dropping variance components to achieve model identification

A second step toward model reduction could be to remove variance components to achieve model identification. Starting with the smallest variance component (or a set of them) this step can be repeated until the PCA no longer suggests overidentification. For the present case, variance components 5 and 7 for `SubjID` and 1 and 4 for `ItemID` are estimated with zero or close to zero values. We refit the LMM without these variance components. The PCA for this LMM estimates 12 non-zero variance components.

### 3.1.4. Dropping non-significant variance components

In the third step, we attempt to simplify the random-effects structure of the identified LMM with likelihood ratio tests. For example, the two smallest variance components account for less than 1% of the variance. We iteratively remove variance components, starting with dropping the highest-order interaction term `SPC`. Moving on to tests of two-factor interactions, we end up with an LMM comprising only varying intercepts for subject and items and the item-related variance component for `P`. Looking back to the maximally identified LMM, we see that these are exactly the three variance components with clearly larger standard-deviation estimates ( $> 249$ ) compared to the other standard-deviation estimates ( $> 64$ ). There is no significant loss of goodness of fit when we remove nine variance components identified this way;  $\chi_9^2 = 11.1$ ,  $p = .27$ . However, removal of any of three remaining variance components significantly reduces the goodness of fit.

### 3.1.5. Extending the reduced LMM with a correlation parameter

Inclusion of the correlation parameter between item-related intercept and the precedence effect (`P`) for this model significantly improves the goodness of fit with the correlation parameter estimated at  $-0.69$ ;  $\chi_1^2 = 16.3$ ,  $p < .01$ . Thus, there is evidence for reliable differences between items in the precedence effect. The variance components and correlation parameter for this final LMM are displayed in Figure 2.†

†Incidentally, although we consider it questionable to compare non-identified and identified models with an LRT, we want to mention that there is no significant difference in goodness of fit between the final LMM and the maximal model we started with. The final number of principal components suggested by the final model is actually smaller than suggested by the initial PCA of maximal model.

### 3.1.6. Summary

In our opinion, the final model we settled on is the *optimal* LMM for the data of this experiment. To summarize our general strategy: (1) we started with a maximal model; (2) then, we fit a zero-correlation model; (3) next, we removed variance components until the likelihood ratio test showed no further improvement; and (4) finally, we added correlation parameters for the remaining variance components. Principal components analysis was used throughout to check the dimensionality for the respective intermediate models. This approach worked quite well in the present case. Indeed, we also reanalyzed three additional experiments reported in the supplement of Barr et al. (2013). As documented in the `RePsychLing` package accompanying the present article, in each case, the maximal LMM was too complex for the information provided by the experimental data. In each case, the data supported only a very sparse random-effects structure beyond varying intercepts for subjects and items. Fortunately and interestingly, none of the analyses impacted the statistical inference about fixed effects in these experiments. Obviously, this cannot be ruled out in general. If authors adhere to a strict criterion for significance, such as  $p < .05$  suitably adjusted for multiple comparisons, there is always a chance that a t-value will fall above or below the criterion across different versions of an LMM.

### 3.2. An alternative analysis of Kronmüller and Barr (2007) using a Bayesian LMM

We also show that similar conclusions can be reached if we fit a Bayesian linear mixed model (Gelman et al., 2014) instead of the frequentist model discussed above using ‘`lme4`’. Presenting the Bayesian estimates corresponding to the maximal linear mixed model presented above provides an independent validation of the conclusion that a simpler model has better motivation.

We fit a linear mixed model to the Kronmüller and Barr data using `rstan` (Stan Development Team, 2018). In a Bayesian linear mixed model, all parameters have a prior distribution defined over them; this is in contrast to the frequentist approach, where each parameter is assumed to have a fixed but unknown value. Defining a prior distribution over each parameter expresses the researcher’s existing knowledge about possible values that the parameter can take, before any new data is considered. For example, in the Bayesian formulation, an `lme4` style model specification such as

```
Y ~ 1 + A + (1|Subject)
```

we could specify that our prior belief about the parameter, call it  $\beta_1$ , expressing an effect of A is that it has a normal distribution with mean 0 and some large variance  $\sigma^2$ . We can write this as

$$\beta_1 \sim \text{Normal}(0, \sigma^2). \tag{1}$$

Such a prior expresses the belief that, in the absence of any new data, the mean is assumed to be zero, but the large variance expresses uncertainty about this belief. Using computational methods available in `rstan`, this prior specification can be combined with the data to derive a posterior distribution for each parameter, including the random

effects variance components and the correlations between variance components. The posterior distribution of each parameter is effectively a compromise between the prior and the data, and expresses our revised belief about the parameter’s distribution after the data are taken into account. If there is strong evidence from the data that the mean of its distribution is different from zero, the mean of the posterior distribution will reflect this. If there is only weak or no evidence from the data—either due to there being too little data or because the mean from the data is near zero—that the parameter has a mean different from zero, then the prior mean of zero will dominate in determining the parameter’s posterior distribution.

The end-product of a Bayesian linear mixed model is always a posterior distribution for each parameter in the model. Thus, we can plot the 95% credible interval for each parameter; this interval tells us the range over which we can be 95% certain that the true value of the parameter lies, given the data. Contrast this with the 95% confidence interval (CI), which represents one of hypothetically computed CIs over repeated experiments, where 95% of those hypothetical CIs would contain the true value of the parameter. Note, however, that for relatively large data-sets such as the present one, the credible interval and confidence interval for the fixed effects will generally be identical (see Figure 1).

We fit a linear mixed model to the Kronmüller and Barr data with normal priors on the fixed effects parameters, and a so-called `lkj` prior on the correlation matrices of the subject and item random effects (Stan Development Team, 2016). The `lkj` prior assumes that the correlations are zero (with some uncertainty associated with this belief); if there is evidence in the data for a non-zero correlation, the posterior distribution of the correlation parameter will be shifted away from zero. For the standard deviations, we defined a uniform prior with a bound 0. This prior expresses that we have no strong beliefs about the standard deviation, but we know that it cannot be less than 0 (our analysis does not depend on using this prior). For a detailed specification of the model, see the `RePsychLing` package.

The posterior distributions of all parameters for the `maximal` model, along with their 95% credible intervals, are shown in Figures 1, 2, 3, and 4. The Bayesian analysis shows two important things. First, the estimates of the fixed effects in the `lme4` model and the Bayesian model are nearly identical. This shows that the “maximal” LMM fit using `lme4` is essentially equivalent to fitting a Bayesian LMM with regularizing priors of the sort described above. Second, the relevant variance component parameters that were identified above using principal components analysis (PCA) and likelihood ratio tests (LRTs) are exactly the parameters that clearly dominate in the Bayesian analysis; see Figures 2, 3, and 4. Specifically, any variance component excluded in the `lme4`-based analysis using PCA and LRTs has, in the Bayesian analysis, a posterior credible interval that includes zero; and any correlation parameter excluded in the `lme4`-based analysis has, in the Bayesian model, a credible interval that spans zero. In other words, when we approach the analysis from the perspective of Bayesian modeling, we also find that there is no evidence in the data that the relevant parameters have values that are different from zero. These parameters should be excluded from the model on grounds of parsimony. For comparison, we also present Bayesian estimates of the final model with a reduced number of variance components (Figures 1-4).

### 3.3. *Reanalysis of Kliegl et al. (2015)*

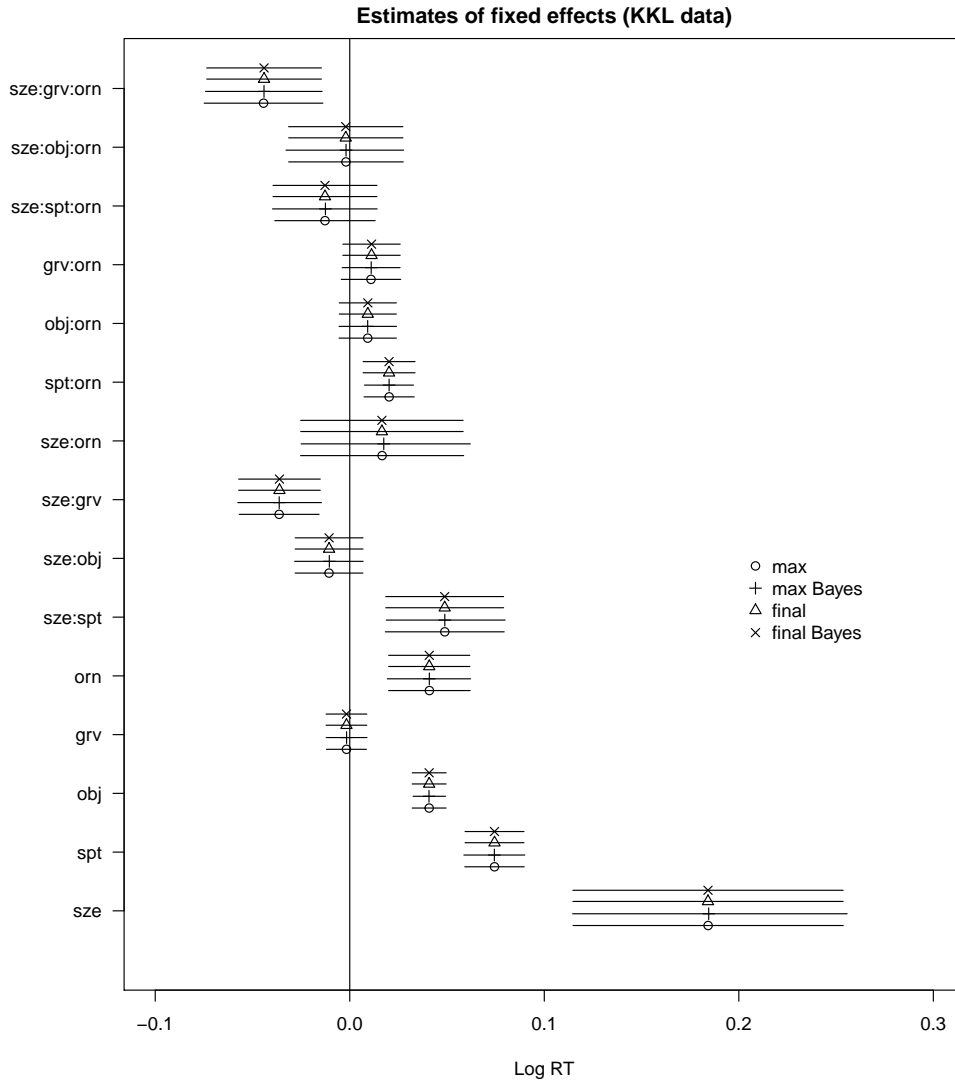
As a second demonstration that linear mixed models with a maximal random-effect structure may be asking too much, we re-analyze data from a visual-attention experiment (Kliegl et al., 2015), following up a published experiment (Kliegl et al., 2011) with an (unfortunately) overidentified LMM, as shown in the vignette for the KWDYZ data in the `RePsychLing` package. The experiment shows that validly cued targets on a monitor are detected faster than invalidly cued ones (i.e., spatial cueing effect; Posner (1980)) and that targets presented at the opposite end of a rectangle at which the cue had occurred are detected faster than targets presented at a different rectangle but with the same physical distance (object-based effect; Egly et al. (1994)). Different from earlier research, the two rectangles were not only presented in cardinal orientation (i.e., in horizontal or vertical orientation), but also diagonally (45° left or 45° right). This manipulation afforded a follow-up of a hypothesis that attention can be shifted faster diagonally than vertically or horizontally across the screen (Kliegl et al., 2011; Zhou et al., 2006). Finally, data are from two groups of subjects, one group had to detect small targets and the other large targets. The experiment is a follow-up to Kliegl et al. (2011) who used only small targets and only cardinal orientations for rectangles. For the interpretation of the fixed effects, we refer to Kliegl et al. (2015). Again, the different model specifications reported in this section were of no consequence for the significance or interpretation of fixed effects, but they led to inappropriate conclusions about the correlations between variance components. We focus here on exploring the random-effect structure for these data.

#### 3.3.1. *Maximal linear mixed model*

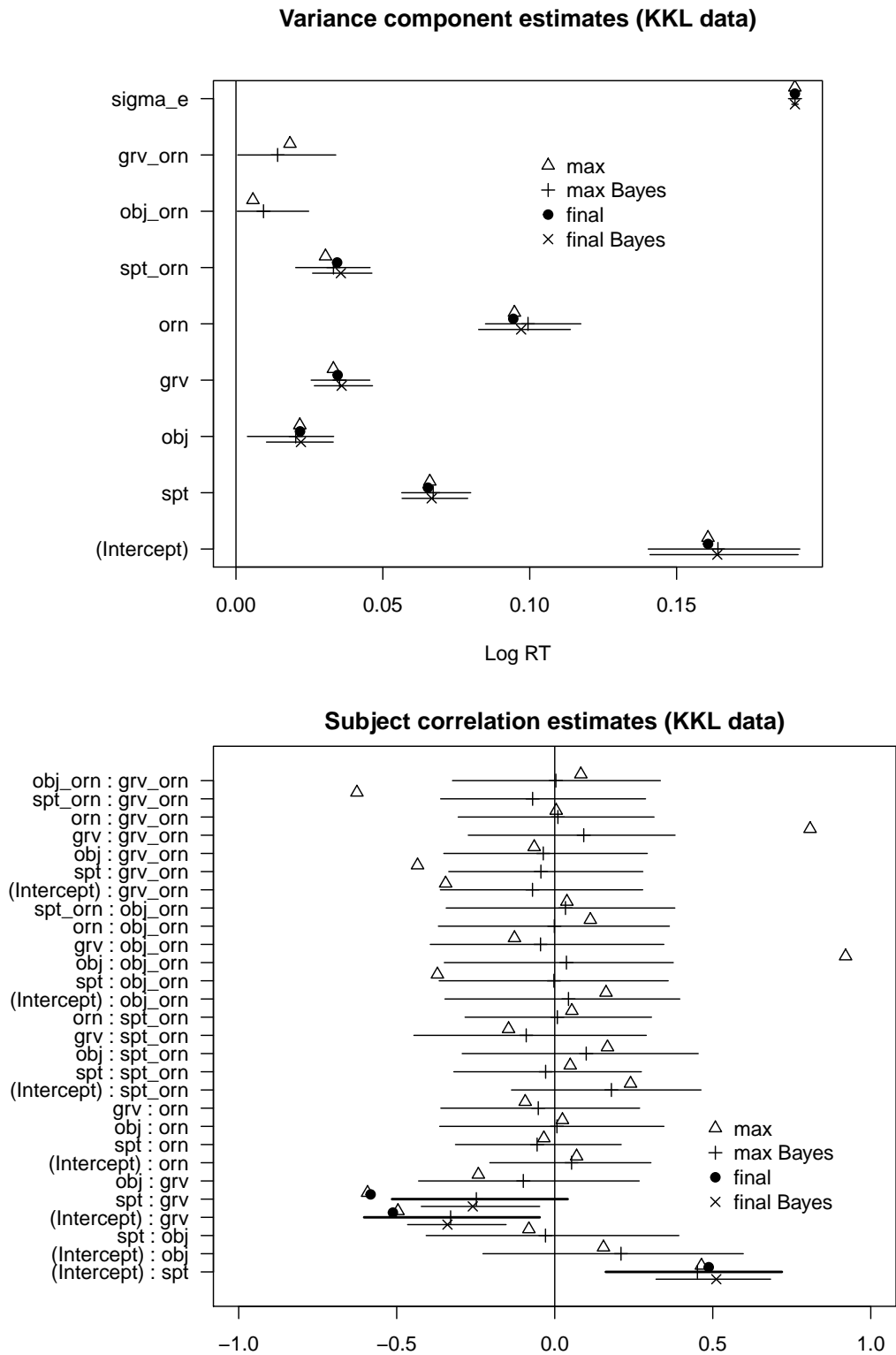
We start with the maximal linear mixed model including all possible variance components and correlation parameters associated with the four within-subject contrasts in the random-effects structure. Note that there are no interactions between the three contrasts associated with the four levels of the cue-target relation factor. Also, as factor size was manipulated between subjects, this contrast does not appear in the random-effect structure. Thus, the random-effect structure comprises eight variance components (i.e., the intercept estimating the grand mean of log reaction time, the three contrasts for the four types of cue-target relation, the contrast for the orientation factor, and three interactions) and 28 correlation parameters ( $8 \times 7/2$ )—also a very complex model. The maximal model converges with a warning:

```
maxfun < 10 *length(par)^2 is not recommended
```

This suggests that we may be asking too much of these data. Nevertheless, at a first glance, model parameters look reasonable. As shown in Figure 6, none of the eight variance components are estimated at zero and none of the 28 correlation parameters are at the boundary (i.e., none assume values of +1 or -1). The PCA, however, indicates that the maximal model is overparameterized: two dimensions contribute 0% variance explained.



**Fig. 5.** The estimates and 95% confidence intervals for the fixed effects in the maximal and final models of the Kliegl et al. 2015 data. Also shown are estimates and 95% credible intervals from maximal and final Bayesian hierarchical linear models.



**Fig. 6.** Top: The standard deviation estimates from the maximal and final models of the KKL data, and estimates and 95% credible intervals of the maximal and final Bayesian hierarchical models. Bottom: Correlations of subject random effects in the maximal and final models of the KKL data, along with estimates and credible intervals from the maximal and final Bayesian model.



### 3.3.2. Zero-correlation-parameter linear mixed models

The problem of overidentification persists in the zero-correlation parameter model (LMM). In the PCA, we still have only seven of eight non-zero components and one of them accounts for less than 1% of the variance. Thus, the LMM still is too complex for the information contained in the data of this experiment.

### 3.3.3. Dropping variance components to achieve model identification

The estimates of variance components suggest that there is very little reliable variance associated with the interaction between object and orientation contrasts and for the interaction between gravitation and orientation. Dropping these two variance components from the model and refitting leads to an identified LMM. Thus, the data of this experiment support six variance components, in agreement with the initial PCA of the maximal model.

### 3.3.4. Testing non-significant variance components

Given an identified LMM, we test whether removal of any of the remaining variance components reduces the goodness of fit in an LRT. It turns out that all of these variance components are reliable. So we keep them in the model.

### 3.3.5. Extending the reduced LMM with correlation parameters

Having arrived at an identified reduced LMM, we expand the LMM and check whether there are significant correlation parameters.‡ This model is also supported by the data: There is no evidence of degeneration. Moreover, the model fits significantly better than the zero-correlation parameter model;  $\chi^2_{15} = 50, p < .01$ . Thus, we would consider this LMM as an acceptable model. The results are documented in Figures 5 and 6.

### 3.3.6. Pruning low correlation parameters

The significant increase in goodness of fit when going from the reduced zero-correlation parameter model to the extended LMM suggests that there is significant information associated with the ensemble of correlation parameters. Nevertheless, the object and orientation effects and the interaction between spatial and orientation effects are only weakly correlated with the mean as well as with spatial and gravitation effects. So we remove these correlation parameters from the model. There is no loss of goodness of fit associated with dropping most of the correlation parameters;  $\chi^2_{12} = 8.4, p = .75$ .

### 3.3.7. Summary

The data from this experiment were a follow-up to an experiment reported by Kliegl et al. (2011). The statistical inferences in that article, especially also with respect to

‡There is a slight risk that we removed a variance component that would have been significant with correlation parameters in the LMM, but we found no evidence for this in additional analyses.

correlation parameters, were based on a maximal LMM. A reanalysis along the strategy described here revealed an overparameterization, involving a negative correlation between spatial and attraction effect and a positive correlation between mean and spatial effect. The reanalysis of those early data is also part of the `RePsychLing` package accompanying the present article. The theoretically important negative and positive correlation parameters were replicated with the present experiment in the absence of problems with model complexity (see Kliegl et al., 2015 for further discussion.)

### 3.3.8. *An analysis of Kliegl et al. (2015) using a Bayesian LMM*

We also fit a maximal Bayesian linear mixed model using `rstan`. As in the analysis of the Kronmüller and Barr (KB) data, this also showed that the same variance components whose credible intervals do not include zero were the ones that the iterative procedure identified as suitable for inclusion. As in the KB data, notice that the means of the correlation estimates in the Bayesian model tend to be closer to zero than those from `lme4`. This is because the prior on the correlation matrix has most of its probability mass around zero. If the sample size is small, the prior will dominate in determining the posterior distribution of the correlations; but if there is sufficient data, and if a non-zero correlation is truly present, the mean of the posterior distribution could be different from zero. We see this in the case of three correlation parameters (Figure 6). For comparison, we also show the estimates from a Bayesian model corresponding to the final LMM chosen in the ‘`lme4`’ analysis presented above.

In sum, the Bayesian analysis independently validates the conclusions based on the PCA-based approach described above.

## 4. Discussion

An important goal in statistical analysis of empirical data is the avoidance of overfitting. Any given data-set can tolerate only a limited number of parameters. Mixed-effects modeling is no exception. In the statistical literature on fitting mixed-effects modeling (see, e.g., Pinheiro and Bates, 2000, Galecki and Burzykowski, 2013, Bates et al., 2015a), the approach taken is one in which variance components are added to the model step by step, typically driven by theoretical considerations. The recommendation of Barr et al. (2013) to fit `maximal` models with all possible random effect components included comes from a very different tradition in which statistics is used to provide a verdict on significance in factorial designs. The authors based their recommendation on a simulation study indicating that anti-conservative results were best avoided by fitting models with as rich a random effects structure as possible.

It is indeed important to make sure that the proper variance components are included in the mixed model. Failure to do so may result in anti-conservative conclusions. However, the advice to “keep it maximal” often creates hopelessly over-specified random effects because the number of correlation parameters to estimate rises quickly with the dimension of the random-effects vectors. The information in the data may not be sufficient to support estimations of such complex models and may result in singular covariance matrices, even when the LMM is identifiable in principle. In this case, we need to replace the complex LMM specification by a more parsimonious one.

With an iterative reduction of the complexity of a degenerate maximal model, one can obtain a model in which the estimated parameters are in line with the information present in the data. We proposed (1) to use PCA to determine the dimensionality of the variance-covariance matrix of the random-effect structure, (2) to initially constrain correlation parameters to zero, especially when an initial attempt to fit a maximal model does not converge, and (3) to drop non-significant variance components and their associated correlation parameters from the model. Each of these reductions may lead to a significant loss in goodness of fit according to LRTs for nested models, in which case this clarifies that the parameter is actually well-supported by information in the data.

Importantly, failure to converge is not due to defects of the estimation algorithm, but is a straightforward consequence of attempting to fit a model that is too complex to be properly supported by the data. We have presented examples showing that the problem of overspecification may arise irrespective of whether estimation is based on maximum likelihood or on Bayesian hierarchical modeling. Furthermore, even under convergence, overparameterization may lead to uninterpretable models, which is why we developed a diagnostic tool for detecting overparameterization.

What one typically finds for overspecified, degenerate models — degenerate because they afford no predictive power beyond an identified model for the same data — is that the presence of superfluous variance components has minute effects on the estimates of the variability of the fixed-effects estimates. They may occasionally affect standard errors in the decimals, pushing a  $p$ -value below or lifting one above some supposedly ‘critical’ level of  $\alpha$ . This should not make a difference as far as conclusions regarding the fixed effects parameters are concerned (Bates et al., 2015a). In fact, comparing parsimonious models with the maximal models discussed by Barr et al. (see the `RePsychLing` package for R for full details on the analyses), there is not a single instance where conclusions about the fixed-effect predictors diverge. Thus, for these real data sets, it is not necessary to aim for maximality when the interest is in a confirmatory analysis of factorial contrasts.

If for some reason it is critical to establish the reliability of a specific variance component or correlation parameter, the most promising approach, where feasible, is to collect more data. Beyond that we must mind the Sunset Salvo (Tukey, 1986): “The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data” (p.74–75).

What then about the simulation studies on which Barr et al. base their recommendations for maximality? Several issues arise here, which all relate to how representative these simulations are with respect to real data sets. First, the simulations implement a factorial contrast that is atypically large compared to what is found in natural data. Second, and more importantly, the correlations in the random effects structure range from  $-0.8$  to  $+0.8$ . Such large correlation parameters are indicative of overparameterization. They hardly ever represent true correlations in the population. As a consequence, these simulations do not provide a solid foundation for recommendations about how to fit mixed-effects models to empirical data.

In summary, maximal models are not necessary to protect against anti-conservative conclusions. This protection is fully provided by comprehensive models that are guided by realistic expectations about the complexity that the data can support. In statistics,

as elsewhere in science, parsimony is a virtue, not a vice.

## References

- Baayen, R. H. (2008) *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge, U.K.: Cambridge University Press.
- Baayen, R. H., Davidson, D. J. and Bates, D. (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, **59**, 390–412.
- Barr, D. J., Levy, R., Scheepers, C. and Tily, H. J. (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, **68**, 255–278.
- Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015a) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, in press. URL: <http://arxiv.org/abs/1406.5823>.
- Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015b) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.
- Clark, H. (1973) The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, **12**, 335–359.
- Egley, R., Driver, J. and Rafal, R. D. (1994) Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, **123**, 161–177.
- Forster, K. and Dickinson, R. (1976) More on the language-as-fixed effect: Monte-Carlo estimates of error rates for  $F_1$ ,  $F_2$ ,  $F'$ , and  $minF'$ . *Journal of Verbal Learning and Verbal Behavior*, **15**, 135–142.
- Galecki, A. and Burzykowski, T. (2013) *Linear mixed-effects models using R. A step-by-step approach*. New York: Springer.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014) *Bayesian Data Analysis*. Chapman and Hall/CRC, third edn.
- Judd, C. M., Westfall, J. and Kenny, D. A. (2012) Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, **103**, 54–69.
- Kliegl, R. (2007) Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, and Reichle (2007). *Journal of experimental psychology: general*, **136**, 530–537.
- Kliegl, R., Kuschela, J. and Laubrock, J. (2015) Object orientation and target size modulate the speed of visual attention. Unpublished manuscript, University of Potsdam.

- Kliegl, R., Risse, S. and Laubrock, J. (2007) Preview benefit and parafoveal-on-foveal effects from word  $n+2$ . *Journal of Experimental Psychology: Human Perception and Performance*, **33**, 1250–1255.
- Kliegl, R., Wei, P., Dambacher, M., Yan, M. and Zhou, X. (2011) Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, **1**, 1–12.
- Kronmüller, E. and Barr, D. J. (2007) Perspective-free pragmatics: Broken precedents and the recovery-from-preemption hypothesis. *Journal of Memory and Language*, **56**, 436–455.
- Oberauer, K. and Kliegl, R. (2006) A formal model of capacity limits in working memory. *Journal of Memory and Language*, **55**, 601–626.
- Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-effects models in S and S-PLUS*. Statistics and Computing. New York: Springer.
- Posner, M. I. (1980) Orienting of attention. *Quarterly Journal of Experimental Psychology*, **32**, 3–25.
- Stan Development Team (2016) *Stan Modeling Language Users Guide and Reference Manual, Version 2.12.0*. URL: <http://mc-stan.org/>.
- (2018) RStan: the R interface to Stan. URL: <http://mc-stan.org/>. R package version 2.17.3.
- Tukey, J. W. (1986) Sunset salvo. *The American Statistician*, **40**, 72–76.
- Vasishth, S. (2003) *Working memory in sentence comprehension: Processing Hindi center embeddings*. New York: Garland Press. Published in the Garland series Outstanding Dissertations in Linguistics, edited by Laurence Horn.
- Vasishth, S. and Lewis, R. L. (2006) Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, **82**, 767–794.
- Westfall, J., Kenny, D. A. and Judd, C. M. (2014) Replicating studies in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, **143**, 2020–2045.
- Zhou, X., Chu, H., Li, X. and Zhan, Y. (2006) Center of mass attracts attention. *Neuroreport*, **17**, 85–88.