Abstract

This paper provides an introduction to mixed-effects models for the analysis of repeated measurement data with subjects and items as crossed random effects. A worked-out example of how to use recent software for mixed-effects modeling is provided. Simulation studies illustrate the advantages offered by mixed-effects analyses compared to traditional analyses based on quasi-F tests, by-subjects analyses, combined by-subjects and by-items analyses, and random regression. Applications and possibilities across a range of domains of inquiry are discussed.

Keywords: Mixed-effects models; Crossed random effects; Quasi-F; By-item; By-subject

Psycholinguists and other cognitive psychologists use convenience samples for their experiments, often based on participants within the local university community. When analyzing the data from these experiments, participants are treated as random variables, because the interest of most studies is not about experimental effects present only in the individuals who participated in the experiment, but rather in effects present in language users everywhere, either within the language studied, or human language users in general. The differences between individuals due to genetic, developmental, environmental, social, political, or chance factors are modeled jointly by means of a participant random effect.

A similar logic applies to linguistic materials. Psycholinguists construct materials for the tasks that they employ by a variety of means, but most importantly, most materials in a single experiment do not exhaust all possible syllables, words, or sentences that could be found in a given language, and most choices of language to investigate do not exhaust the possible languages that an experimenter could investigate. In fact, two core principles of the structure of language, the arbitrary (and hence statistical) association between sound and meaning and the unbounded combination of finite lexical items, guarantee that a great many language materials must be a sample, rather than an exhaustive list. The space of possible words, and the space of possible sentences, is simply too large to be modeled by any other means. Just as we model human participants as random variables, we have to model factors characterizing their speech as random variables as well.
of modeling heteroskedasticity and non-spherical error variance (for either participants or items). Methods for estimating linear mixed effect models have addressed each of these concerns, and offer a better approach than univariate ANOVA or ordinary least squares regression.

In what follows, we first introduce the concepts and formalism of mixed effects modeling.

### Mixed effects model concepts and formalism

The concepts involved in a linear mixed effects model will be introduced by tracing the data analysis path of a simple example. Assume an example data set with three participants $s_1$, $s_2$ and $s_3$ who each saw three items $w_1$, $w_2$, $w_3$ in a priming lexical decision task under both short and long SOA conditions. The design, the RTs and their constituent fixed and random effects components are shown in Table 1.

This table is divided into three sections. The leftmost section lists subjects, items, the two levels of the SOA factor, and the reaction times for each combination of subject, item and SOA. This section represents the data available to the analyst. The remaining sections of the table list the effects of SOA and the properties of the subjects and items that underly the RTs. Of these remaining sections, the middle section lists the fixed effects: the intercept (which is the same for all observations) and the effect of SOA (a 19 ms processing advantage for the short SOA condition). The right section of the table lists the random effects in the model. The first column in this section lists by-item adjustments to the intercept, and the second column lists by-subject adjustments to the intercept. The third column lists by-subject adjustments to the effect of SOA. For instance, for the first subject the effect of a short SOA is attenuated by 11 ms. The final column lists the by-observation noise. Note that in this example we did not include by-item adjustments to SOA, even though we could have done so. In the terminology of mixed effects modeling, this data set is characterized by random intercepts for both subject and item, and by by-subject random slopes (but not by-item random slopes) for SOA.

Formally, this dataset is summarized in (1).

$$y_{ij} = \mathbf{X}_{ij} \beta + S_i s_i + W_j w_j + c_{ij}$$ (1)

The vector $y_{ij}$ represents the responses of subject $i$ to item $j$. In the present example, each of the vectors $y_{ij}$ comprises two response latencies, one for the short and one for the long SOA. In (1), $\mathbf{X}_{ij}$ is the design matrix, consisting of an initial column of ones and followed by columns representing factor contrasts and covariates. For the present example, the design matrix for each subject-item combination has the simple form
Table 1
Example data set with random intercepts for subject and item, and random slopes for subject

<table>
<thead>
<tr>
<th>Subj</th>
<th>Item</th>
<th>SOA</th>
<th>RT</th>
<th>Fixed</th>
<th>ItemInt</th>
<th>Random</th>
<th>Res</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>w1</td>
<td>Long</td>
<td>466</td>
<td>522.2</td>
<td>0</td>
<td>-28.3</td>
<td>-26.2</td>
</tr>
<tr>
<td>s1</td>
<td>w2</td>
<td>Long</td>
<td>520</td>
<td>522.2</td>
<td>0</td>
<td>14.2</td>
<td>-26.2</td>
</tr>
<tr>
<td>s1</td>
<td>w3</td>
<td>Long</td>
<td>502</td>
<td>522.2</td>
<td>0</td>
<td>14.1</td>
<td>-26.2</td>
</tr>
<tr>
<td>s1</td>
<td>w1</td>
<td>Short</td>
<td>475</td>
<td>522.2</td>
<td>-19</td>
<td>-28.3</td>
<td>-26.2</td>
</tr>
<tr>
<td>s1</td>
<td>w2</td>
<td>Short</td>
<td>494</td>
<td>522.2</td>
<td>-19</td>
<td>14.2</td>
<td>-26.2</td>
</tr>
<tr>
<td>s1</td>
<td>w3</td>
<td>Short</td>
<td>490</td>
<td>522.2</td>
<td>-19</td>
<td>14.1</td>
<td>-26.2</td>
</tr>
<tr>
<td>s2</td>
<td>w1</td>
<td>Long</td>
<td>516</td>
<td>522.2</td>
<td>0</td>
<td>-28.3</td>
<td>29.7</td>
</tr>
<tr>
<td>s2</td>
<td>w2</td>
<td>Long</td>
<td>566</td>
<td>522.2</td>
<td>0</td>
<td>14.2</td>
<td>29.7</td>
</tr>
<tr>
<td>s2</td>
<td>w3</td>
<td>Long</td>
<td>577</td>
<td>522.2</td>
<td>0</td>
<td>14.1</td>
<td>29.7</td>
</tr>
<tr>
<td>s2</td>
<td>w1</td>
<td>Short</td>
<td>491</td>
<td>522.2</td>
<td>-19</td>
<td>-28.3</td>
<td>29.7</td>
</tr>
<tr>
<td>s2</td>
<td>w2</td>
<td>Short</td>
<td>544</td>
<td>522.2</td>
<td>-19</td>
<td>14.2</td>
<td>29.7</td>
</tr>
<tr>
<td>s2</td>
<td>w3</td>
<td>Short</td>
<td>526</td>
<td>522.2</td>
<td>-19</td>
<td>14.1</td>
<td>29.7</td>
</tr>
<tr>
<td>s3</td>
<td>w1</td>
<td>Long</td>
<td>484</td>
<td>522.2</td>
<td>0</td>
<td>-28.3</td>
<td>-3.5</td>
</tr>
<tr>
<td>s3</td>
<td>w2</td>
<td>Long</td>
<td>529</td>
<td>522.2</td>
<td>0</td>
<td>14.2</td>
<td>-3.5</td>
</tr>
<tr>
<td>s3</td>
<td>w3</td>
<td>Long</td>
<td>539</td>
<td>522.2</td>
<td>0</td>
<td>14.1</td>
<td>-3.5</td>
</tr>
<tr>
<td>s3</td>
<td>w1</td>
<td>Short</td>
<td>470</td>
<td>522.2</td>
<td>-19</td>
<td>-28.3</td>
<td>-3.5</td>
</tr>
<tr>
<td>s3</td>
<td>w2</td>
<td>Short</td>
<td>511</td>
<td>522.2</td>
<td>-19</td>
<td>14.2</td>
<td>-3.5</td>
</tr>
<tr>
<td>s3</td>
<td>w3</td>
<td>Short</td>
<td>528</td>
<td>522.2</td>
<td>-19</td>
<td>14.1</td>
<td>-3.5</td>
</tr>
</tbody>
</table>

The first four columns of this table constitute the data normally available to the researcher. The remaining columns parse the RTs into the contributions from the fixed and random effects. Int: intercept, SOA: contrast effect for SOA; ItemInt: by-item adjustments to the intercept; SubInt: by-subject adjustments to the intercept; SubSOA: by-subject adjustments to the slope of the SOA contrast; Res: residual noise.

The next two terms in Eq. (1) serve to make the model’s predictions more precise for the subjects and items actually examined in the experiment. First consider the random effects structure for Subject. The \( S_1 \) matrix is again a full copy of the \( X_0 \) matrix. It is multiplied with a vector specifying for subject \( i \) the adjustments that are required for this subject to the intercept and to the SOA contrast coefficient. For the first subject in Table 1,

\[ S_{i_1} s_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} -26.2 \\ 11.0 \end{pmatrix} = \begin{pmatrix} -26.2 \\ -15.2 \end{pmatrix}, \]

which tells us, first, that for this subject the intercept has to be adjusted downwards by 26.2 ms for both the long and the short SOA (the subject is a fast responder) and second, that, in the short SOA condition the effect of SOA for this subject is attenuated by 11.0 ms. Combined with the adjustment for the intercept that also applies to the short SOA condition, the net outcome for an arbitrary item in the short SOA condition is \(-15.2\) ms for this subject.

Further precision is obtained by bringing the item random effect into the model. The \( W_j \) matrix is again a copy of the design matrix \( X_j \). In the present example, only the first column, the column for the intercept, is retained. This is because in this particular constructed data set the effect of SOA does not vary systematically with item. The vector \( w_1 \), therefore contains one element only for each item \( j \). This element specifies the adjustment made to the population intercept to calibrate the expected values for the specific processing costs associated with this individual item. For item 1 in our example, this adjustment is \(-28.3\), indicating that compared to the population average, this particular item elicited...
shorter latencies, for both SOA conditions, across all subjects.

\[
\mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (28.3) = \begin{pmatrix} 28.3 \\ 28.3 \end{pmatrix}
\]

(6)

The model specification (1) has as its last term the vector of residual errors \( \mathbf{\epsilon}_y \), which in our running example has two elements for each combination of subject and item, one error for each SOA.

For subject i, Eq. (1) formalizes the following vector of sums,

\[
y_i = y_{ij} = X_j \beta + S_{si} + W_{w_j} + \epsilon_{ij} =
\]

which we can rearrange in the form of a composite intercept, followed by a composite effect of SOA, followed by the residual error.

\[
y_i = \begin{pmatrix} (28.2 - 26.2 - 28.3) + (0 + 0) \\ (28.2 - 26.2 + 14.2) + (0 + 0) \\ (28.2 - 26.2 - 28.3) + (0 + 0) \\ (28.2 - 26.2 + 14.2) + (0 + 0) \end{pmatrix} + \begin{pmatrix} 28.3 \\ 14.2 \\ -28.3 \\ 14.2 \end{pmatrix} + \begin{pmatrix} 2 \end{pmatrix}
\]

(7)

Because the random slopes and intercept are pairwise tied to the same observational units, they may be correlated. For our data, \( \rho_{\text{sub}, \text{mis}} = -0.71 \). These four random effects parameters complete the specification of the quantitative structure of our dataset. We can now present the full formal specification of the corresponding mixed-effects model,

\[
y = \mathbf{X} \beta + \mathbf{Z} b + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad b \sim \mathcal{N}(0, \sigma^2 \mathbf{\Sigma}), \quad \mathbf{b} \perp \epsilon,
\]

(10)

where \( \mathbf{\Sigma} \) is the relative variance-covariance matrix for the random effects. The symbol \( \perp \) indicates independence of random variables and \( \mathcal{N} \) denotes the multivariate normal (Gaussian) distribution. We say that matrix \( \mathbf{\Sigma} \) is the relative variance-covariance matrix of the random effects in the sense that it is the variance of \( \mathbf{b} \) relative to \( \sigma^2 \), the scalar variance of the per-observation noise term \( \epsilon \). The variance-covariance specification of the model is an important tool to capture non-independence (asphericity) between observations.

Hypotheses about the structure of the variance-covariance matrix can be tested by means of likelihood ratio tests. Thus, we can formally test whether a random effect for items is required and that the presence of the parameter \( \sigma_i \) in the model specification is actually justified. Similarly, we can inquire whether a parameter for the covariance of the by-subject slopes and intercepts contributes significantly to the model’s goodness of fit. We note that in this approach it is an empirical question whether random effects for item or subject are required in the model.

When a mixed-effects model is fitted to a data set, its set of estimated parameters includes the coefficients for the
fixed effects on the one hand, and the standard deviations and correlations for the random effects on the other hand. The individual values of the adjustments made to intercepts and slopes are calculated once the random-effects parameters have been estimated. Formally, these adjustments, referenced as Best Linear Unbiased Predictors (or BLUPs), are not parameters of the model.

**Data analysis**

We illustrate mixed-effects modeling with R, an open-source language and environment for statistical computing (R development core team, 2007), freely available at http://cran.r-project.org. The lme4 package (Bates, 2005; Bates & Sarkar, 2007) offers fast and reliable algorithms for parameter estimation (see also West et al., 2007:14) as well as tools for evaluating the model (using Markov chain Monte Carlo sampling, as explained below).

Input data for R should have the structure of the first block in Table 1, together with an initial header line specifying column names. The data for the first subject therefore should be structured as follows, using what is known as the long data format in R (and as the univariate data format in SPSS):

```
<table>
<thead>
<tr>
<th>Subj</th>
<th>Item</th>
<th>SOA</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>s1</td>
<td>w1</td>
<td>475</td>
</tr>
<tr>
<td>2</td>
<td>s1</td>
<td>w2</td>
<td>494</td>
</tr>
<tr>
<td>3</td>
<td>s1</td>
<td>w3</td>
<td>490</td>
</tr>
<tr>
<td>4</td>
<td>s1</td>
<td>w1</td>
<td>466</td>
</tr>
<tr>
<td>5</td>
<td>s1</td>
<td>w2</td>
<td>520</td>
</tr>
<tr>
<td>6</td>
<td>s1</td>
<td>w3</td>
<td>502</td>
</tr>
</tbody>
</table>
```

We load the data, here simply an ASCII text file, into R with

```r
> priming = read.table("ourdata.txt", header = TRUE)
```

SPSS data files (if brought into the long format within SPSS) can be loaded with read.spss and csv tables (in long format) are loaded with read.csv. We fit the model of Eq. (10) to the data with

```r
> priming.lmer = lmer(RT ~ SOA + (1|Item) + (1 + SOA|Subj), data = priming)
```

The dependent variable, RT, appears to the left of the tilde operator (~), which is read as “depends on” or “is a function of”. The main effect of SOA, our fixed effect, is specified to the right of the tilde. The random intercept for Item is specified with (1|Item), which is read as a random effect introducing adjustments to the intercept (denoted by 1) conditional on or grouped by Item. The random effects for Subject are specified as (1+SOA|Subject). This notation indicates, first of all, that we introduce by-subject adjustments to the intercept (again denoted by 1) as well as by-subject adjustments to SOA. In other words, this model includes by-subject and by-item random intercepts, and by-subject random slopes for SOA. This notation also indicates that the variances for the two levels of SOA are allowed to be different. In other words, it models potential by-subject heteroskedasticity with respect to SOA. Finally, this specification includes a parameter estimating the correlation \( \hat{\rho}_{\text{int, soa}} \) of the by-subject random effects for slope and intercept.

A summary of the model is obtained with

```r
> summary(priming.lmer)
```

```
Linear mixed-effects model fit by REML

Formula: RT ~ SOA + (1|Item) + (1 + SOA|Subj)

Data: priming

AIC   BIC logLik ML deviance REML deviance
150.0 155.4  -69.02 151.4 138.0

Random effects:

Groups   Name        Variance Std.Dev Corr

Item     (Intercept) 613.73   24.774
         SOAshort     136.46   11.682 -1.000

Subj     (Intercept) 803.07   28.338
         SOAshort     136.46   11.682 -1.000

Residual                         102.28   10.115

number of obs: 18, groups: Item, 3; Subj, 3

Fixed effects:

                 Estimate Std. Error t value

(Intercept)      522.111 21.992 23.741
SOAshort         -18.889   8.259  -2.287

```

The summary first mentions that the model is fitted using restricted maximum likelihood estimation (REML), a modification of maximum likelihood estimation that is more precise for mixed-effects modeling. Maximum likelihood estimation seeks to find those parameter values that, given the data and our choice of model, make the model’s predicted values most similar to the observed values. Discussion of the technical details of model fitting is beyond the scope of this paper. However, in the Appendix we provide some indication of the kind of issues involved.

The summary proceeds with repeating the model specification, and then lists various measures of goodness of fit. The remainder of the summary contains two subtables, one for the random effects, and one for the fixed effects.

The subtable for the fixed-effects shows that the estimates for slope and the contrast coefficient for SOA are right on target: 522.11 for the intercept (compare 522.2 in Table 1), and -18.89 (compare -19.0). For each coefficient, its standard error and t-statistic are listed.
Turning to the subtable of random effects, we observe that the first column lists the main grouping factors: Item, Subj and the observation noise (Residual). The second column specifies whether the random effect concerns the intercept or a slope. The third column reports the variances, and the fourth column the square roots of these variances, i.e., the corresponding standard deviations. The sample standard deviations calculated above on the basis of Table 1 compare well with the model estimates, as shown in Table 2.

The high correlation of the intercept and slope for the subject random effects (−1.00) indicates that the model has been overparameterized. We first simplify the model by removing the correlation parameter and by assuming homoskedasticity for the subjects with respect to the SOA conditions, as follows:

```R
> print(priming.lmer1, corr = FALSE)
Random effects:
 Groups Name Variance Std.Dev.
 SOA:Subj (Intercept) 34.039 5.8343
 Subj (Intercept) 489.487 22.1243
 Item (Intercept) 625.623 25.0125
 Residual 119.715 10.9414

number of obs: 18, groups: SOA:Subj, 6; Subj, 3; Item, 3
 Fixed effects: 
 Estimate Std. Error t value
 (Intercept) 522.111 19.909 26.23
 SOAshort -18.889 7.021 -2.69
```

(Here and in the examples to follow, we abbreviated the R output.) The variance for the by-subject adjustments for SOA is small, and potentially redundant, so we further simplify to a model with only random intercepts for subject:

```R
> priming.lmer2 = lmer(RT ~ SOA + (1|Item) + (1|Subj), data = priming)
```

In order to verify that this most simple model is justified, we carry out a likelihood ratio test (see, e.g., Pinheiro & Bates, 2000, p. 83) that compares the most specific model `priming.lmer2` (which sets ρ to the specific value of zero and assumes homoskedasticity) with the more general model `priming.lmer` (which does not restrict ρ a priori and explicitly models heteroskedasticity). The likelihood of the more general model (Lg) should be greater than the likelihood of the more specific model (Lr), and hence the likelihood ratio test statistic 2log(Lg/Lr) > 0. If g is the number of parameters for the general model, and s the number of parameters for the restricted model, then the asymptotic distribution of the likelihood ratio test statistic, under the null hypothesis that the restricted model is sufficient, follows a chi-squared distribution with g−s degrees of freedom. In R, the likelihood ratio test is carried out with the `anova` function:

```R
> anova(priming.lmer, priming.lmer2)
Df AIC BIC logLik Chisq Chi Df Pr(>Chisq)
priming.lmer2 4 162.353 165.914 -77.176
priming.lmer 6 163.397 168.740 -75.699 2.557 2 0.2282
```

The value listed under Chisq equals twice the difference between the log-likelihood (listed under logLik) for `priming.lmer` and that of `priming.lmer2`. The degrees of freedom for the chi-squared distribution, 2, is the difference between the number of parameters in the model (listed under Df). It is clear that the removal of the parameter for the correlation together with the parameter for by-subject random slopes for SOA is justified ($X^2(2) = 2.557, p = 0.228$). The summary of the simplified model

```R
> print(priming.lmer2, corr = FALSE)
Random effects:
 Groups Name Variance Std.Dev.
 Item (Intercept) 607.72 24.652
 Subj (Intercept) 499.22 22.343
 Residual 119.715 10.9414

number of obs: 18, groups: Item, 3; Subj, 3
 Fixed effects: 
 Estimate Std. Error t value
 (Intercept) 522.111 19.909 26.23
 SOAshort -18.889 7.021 -2.69
```
lists only random intercepts for subject and item, as desired.

The reader may have noted that summaries for model objects fitted with `lme` list standard errors and t-statistics for the fixed effects, but no p-values. This is not without reason.

With many statistical modeling techniques we can derive exact distributions for certain statistics calculated from the data and use these distributions to perform hypothesis tests on the parameters, or to create confidence intervals or confidence regions for the values of these parameters. The general class of linear models fit by ordinary least squares is the prime example of such a well-behaved class of statistical models for which we can derive exact results, subject to certain assumptions on the distribution of the responses (normal, constant variance and independent disturbance terms). This general paradigm provides many of the standard techniques of modern applied statistics including t-tests and analysis of variance decompositions, as well as confidence intervals based on t-distributions. It is tempting to believe that all statistical techniques should provide such neatly packaged results, but they don’t.

Inferences regarding the fixed-effects parameters are more complicated in a linear mixed-effects model than in a linear model with fixed effects only. In a model with only fixed effects we estimate these parameters and one other parameter which is the variance of the noise that infects each observation and that we assume to be independent and identically distributed (i.i.d.) with a normal (Gaussian) distribution. The initial work by William Gossett (who wrote under the pseudonym of “Student”) on the effect of estimating the variance of the disturbances on the estimates of precision of the sample mean, leading to the t-distribution, and later generalizations by Sir Ronald Fisher, providing the analysis of variance, were turning points in 20th century statistics.

When mixed-effects models were first examined, in that days when the computing tools were considerably less sophisticated than at present, many approximations were used, based on analogy to fixed-effects analysis of variance. For example, variance components were often estimated by calculating certain mean squares and equating the observed mean square to the corresponding expected mean square. There is no underlying objective, such as the log-likelihood or the log-restricted-likelihood, that is being optimized by such estimates. They are simply assumed to be desirable because of the analogy to the results in the analysis of variance. Furthermore, the theoretical derivations and corresponding calculations become formidable in the presence of multiple factors, such as both subject and item, associated with random effects or in the presence of unbalanced data.

Fortunately, it is now possible to evaluate the maximum likelihood or the `REML` estimates of the parameters in mixed-effects models reasonably easily and quickly, even for complex models fit to very large observational data sets. However, the temptation to perform hypothesis tests using t-distributions or F-distributions based on certain approximations of the degrees of freedom in these distributions persists. An exact calculation can be derived for certain models with a comparatively simple structure applied to exactly balanced data sets, such as occur in text books. In real-world studies the data often end up unbalanced, especially in observational studies but even in designed experiments where missing data can and do occur, and the models can be quite complicated. The simple formulas for the degrees of freedom for inferences based on t or F-distributions do not apply in such cases. In fact, the pivotal quantities for such hypothesis tests do not even have t or F-distributions in such cases so trying to determine the “correct” value of the degrees of freedom to apply is meaningless. There are many approximations in use for hypothesis tests in mixed models—the `MIXED` procedure in `SAS` offers 6 different calculations of degrees of freedom in certain tests, each leading to different p-values, but none of them is “correct”.

It is not even obvious how to count the number of parameters in a mixed-effects model. Suppose we have 1000 subjects, each exposed to 200 items chosen from a pool of 10000 potential items. If we model the effect of subject and item as independent random effects we add two variance components to the model. At the estimated parameter values we can evaluate 1000 predictors of the random effects for subject and 10000 predictors of the random effects for item. Did we only add two parameters to the model when we incorporated these 11000 random effects? Or should we say that we added several thousand parameters that are adjusted to help explain the observed variation in the data? It is overly simplistic to say that thousands of random effects amount to only two parameters. However, because of the shrinkage effect in the evaluation of the random effects, each random effect does not represent an independent parameter.

Fortunately, we can avoid this issue of counting parameters or, more generally, the issue of approximating degrees of freedom. Recall that the original purpose of the t and F-distributions is to take into account the imprecision in the estimate of the variance of the random disturbances when formulating inferences regarding the fixed-effects parameters. We can approach this problem in the more general context with Markov chain Monte Carlo (`MCMC`) simulations. In `MCMC` simulations we sample from conditional distributions of parameter subsets in a cycle, thus allowing the variation in one parameter subset, such as the variance of the random disturbances or the variances and covariances of random effects, to be reflected in the variation of other parameter subsets, such as the fixed effects. This is what the t and F-distributions accomplish in the case of mod-
els with fixed-effects only. Crucially, the \texttt{mcmc} technique applies to more general models and to data sets with arbitrary structure.

Informally, we can conceive of Markov chain Monte Carlo (MCMC) sampling from the posterior distribution of the parameters (see, e.g., Andrieu, de Freitas, Doucet, & Jordan, 2003, for a general introduction to MCMC) as a random walk in parameter space. Each mixed effects model is associated with a parameter vector, which can be divided into three subsets,

1. the variance, $\sigma^2$, of the per-observation noise term,
2. the parameters that determine the variance-covariance matrix of the random effects, and
3. the random effects $\mathbf{b}$ and the fixed effects $\mathbf{\beta}$.

Conditional on the other two subsets and on the data, we can sample directly from the posterior distribution of the remaining subset. For the first subset we sample from a chi-squared distribution conditional on the current residuals. The prior for the variances and covariances of the random effects is chosen so that for the second subset we sample from a Wishart distribution. Finally, conditional on the first two subsets and on the data the sampling for the third subset is from a multivariate normal distribution. The details are less important than the fact that these are well-accepted “non-informative” priors for these parameters. Starting from the \texttt{reml} estimates of the parameters in the model we cycle through these steps many times to generate a sample from the posterior distribution of the parameters. The \texttt{mcmcsamp} function produces such a sample, for which we plot the estimated densities on a log scale.

```r
> mcmc = mcmcsamp(priming.lmer2, n = 50000)
> densityplot(mcmc, plot.points = FALSE)
```

The resulting plot is shown in Fig. 1. We can see that the posterior density of the fixed-effects parameters is reasonably symmetric and close to a normal (Gaussian) dis-

![Fig. 1. Empirical density estimates for the Markov chain Monte Carlo sample for the posterior distribution of the parameters in the model for the priming data with random intercepts only (priming.lmer2). From top to bottom: $\log \sigma^2_{\text{int}}$, $\log \sigma^2_{\text{res}}$, $\log \sigma^2$, $\beta_{\text{SOAshort}}$, $\beta_{\text{Int.}}$.](attachment:image.png)
distribution, which is generally the case for such parameters. After we have checked this we can evaluate p-values from the sample with an ancillary function defined in the languageR package, which takes a fitted model as input and generates by default 10,000 samples from the posterior distribution:

```r
> mcmc = pvals.fnc(priming.lmer2, nsim = 10000)
> mcmc$pfixed
```

|                  | Estimate | MCMCmean | HPD95lower | HPD95upper | pMCMC | Pr(>|t|) |
|------------------|----------|----------|------------|------------|--------|----------|
| (Intercept)      | 522.11   | 521.80   | 435.45     | 616.641    | 0.0012 | 0.0000   |
| SOAshort         | -18.89   | -18.81   | -32.09     | -6.533     | 0.0088 | 0.0035   |

We obtain p-values for only the first two parameters (the fixed effects). The first two columns show that the model estimates and the mean estimate across MCMC samples are highly similar, as expected. The next two columns show the upper and lower 95% highest posterior density intervals (see below). The final two columns show p-values based on the posterior distribution (pMCMC) and on the t-distribution respectively. The degrees of freedom used for the t-distribution by pvals.fnc() is an upper bound: the number of observations minus the number of fixed-effects parameters. As a consequence, p-values calculated with these degrees of freedom will be anti-conservative for small samples.

The distributions of the log-transformed variance parameters are also reasonably symmetric, although some rightward skewing is visible in Fig. 1. Without the log transformation, this skewing would be much more pronounced: The untransformed distributions would not be approximated well by a normal distribution with mean equal to the estimate and standard deviation equal to a standard error. That the distribution of the variance parameters is not symmetric should not come as a surprise. The use of a \(\chi^2\) distribution for a variance estimate is taught in most introductory statistics courses. As Box and Tiao (1992) emphasize, the logarithm of the variance is a more natural scale on which to assume symmetry.

For each of the panels in Fig. 1 we calculate a Bayesian highest posterior density (HPD) confidence interval. For each parameter the HPD interval is constructed from the empirical cumulative distribution function of the sample as the shortest interval for which the difference in the empirical cumulative distribution function values of the endpoints is the nominal probability. In other words, the intervals are calculated to have 95% probability content. There are many such intervals. The HPD intervals are the shortest intervals with the given probability content. Because they are created from a sample, these intervals are not deterministic: taking another sample gives slightly different values. The HPD intervals for the fixed effects in the present example are listed in the output of pvals.fnc(), as illustrated above. The standard 95% confidence intervals for the fixed effects parameters, according to \(\hat{\beta} = t(x/2, v)\delta\), with the upper bound for the degrees of freedom (18 \(–\) 2 = 16) are narrower:

```r
> coefs <- summary(priming.lmer1)$coefs
> coefs[, 1] + qt(0.975, 16) * outer(coefs[, 2], c(-1, 1))
```

For small data sets such as the example data considered here, they give rise to less conservative inferences that may be incorrect and should be avoided.

The HPD intervals for the random effects can be obtained from the mcmc object obtained with pvals.fnc() as follows:

```r
> mcmc$random
```

<table>
<thead>
<tr>
<th></th>
<th>MCMCmean</th>
<th>HPD95lower</th>
<th>HPD95upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>sigma</td>
<td>12.76</td>
<td>7.947</td>
<td>21.55</td>
</tr>
<tr>
<td>Item.(In)</td>
<td>27.54</td>
<td>6.379</td>
<td>140.96</td>
</tr>
<tr>
<td>Subj.(In)</td>
<td>32.62</td>
<td>9.820</td>
<td>138.47</td>
</tr>
</tbody>
</table>

It is worth noting that the variances for the random effects parameters may get close to zero but will never actually be zero. Generally it would not make sense to test a hypothesis of the form \(H_0: \sigma^2 = 0\) versus \(H_1: \sigma^2 > 0\) for these parameters. Neither “Inverting” the HPD interval nor using the empirical cumulative distribution function from the mcmc sam-
ple evaluated at zero works because the value 0 cannot occur in the MCMC sample. Using the estimate of the variance (or the standard deviation) and a standard error to create a z statistic is, in our opinion, nonsense because we know that the distribution of the parameter estimates is not symmetric and does not converge to a normal distribution. We therefore recommend likelihood ratio tests for evaluating whether including a random effects parameter is justified. As illustrated above, we fit a model with and without the variance component and compare the quality of the fits. The likelihood ratio is a reasonable test statistic for the comparison but we note that the “asymptotic” reference distribution of a $\chi^2$ does not apply because the parameter value being tested is on the boundary. Therefore, the $p$-value computed using the $\chi^2$ reference distribution is conservative for variance parameters. For correlation parameters, which can be both positive or negative, this caveat does not apply.

**Key advantages of mixed-effects modeling**

An important new possibility offered by mixed-effects modeling is to bring effects that unfold during the course of an experiment into account, and to consider other potentially relevant covariates as well.

There are several kinds of longitudinal effects that one may wish to consider. First, there are effects of learning or fatigue. In chronometric experiments, for instance, some subjects start out with very short response latencies, but as the experiment progresses, they find that they cannot keep up their fast initial pace, and their latencies progressively increase. Other subjects start out cautiously, and progressively tune in to the task and respond more and more quickly. By means of counterbalancing, adverse effects of learning and fatigue can be neutralized, in the sense that the risk of confounding these effects with critical predictors is reduced. However, the effects themselves are not brought into the statistical model, and consequently experimental noise remains in the data, rendering more difficult the detection of significance for the predictors of interest when subsets of subjects are exposed to the same lists of items.

Second, in chronometric paradigms, the response to a target trial is heavily influenced by how the preceding trials were processed. In lexical decision, for instance, the reaction time to the preceding word in the experiment is one of the best predictors for the target latency, with effect sizes that may exceed that of the word frequency effect. Often, this predictivity extends from the immediately preceding trial to several additional preceding trials. This major source of experimental noise should be brought under statistical control, at the risk of failing to detect otherwise significant effects.

Third, qualitative properties of preceding trials should be brought under statistical control as well. Here, one can think of whether the response to the preceding trial in a lexical decision task was correct or incorrect, whether the preceding item was a word or a nonword, a noun or a verb, and so on.

Fourth, in tasks using long-distance priming, longitudinal effects are manipulated on purpose. Yet the statistical methodology of the past decades allowed priming effects to be evaluated only after averaging over subjects or items. However, the details of how a specific prime was processed by a specific subject may be revealing about how that subject processes the associated target presented later in the experiment.

Because mixed-effects models do not require prior averaging, they offer the possibility of bringing all these kinds of longitudinal effects straightforwardly into the statistical model. In what follows, we illustrate this advantage for a long-distance priming experiment reported in de Vaan, Schreuder, and Baayen (2007). Their lexical decision experiment used long-term priming (with 39 trials intervening between prime and target) to probe budding frequency effects for morphologically complex neologisms. Neologisms were preceded by two kinds of prime, the neologism itself (identity priming) or its base word (base priming). The data are available in the languageR package in the CRAN archives (http://cran.r-project.org, see Baayen, 2008, for further documentation on this package) under the name primingHeidPrevRT. After attaching this data set we fit an initial model with Subject and Word as random effects and priming Condition as fixed-effect factor.

```r
> attach(primingHeidPrevRT)
> print(lmer(RT ~ Condition + (1|Word) + (1|Subject)), corr = FALSE)
```

**Random effects:**

```
Groups     Name Variance     Std.Dev.
---       --- ------- ------- -------
Word      (Intercept)    0.0034119  0.058412
Subject   (Intercept)    0.0408438  0.202098
Residual                   0.0440838  0.209962
```

number of obs: 832, groups: Word, 40; Subject, 26

**Fixed effects:**

```
Estimate  Std. Error   t value
(Intercept)  6.60297    0.04215   156.66
Conditionheid 0.05127    0.01467     2.13
```

The positive contrast coefficient for Condition and $t > 2$ in the summary suggests that long-distance identity priming would lead to significantly longer response latencies compared to base priming.
However, this counterintuitive inhibitory priming effect is no longer significant when the decision latency at the preceding trial \((\text{RT}_\text{min1})\) is brought into the model,

```r
> print(lmer(RT ~ log(RTmin1) + Condition + (1|Word) + (1|Subject)), cor = FALSE)
```

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>(Intercept)</td>
<td>0.0034623</td>
<td>0.058841</td>
</tr>
<tr>
<td>Subject</td>
<td>(Intercept)</td>
<td>0.0334773</td>
<td>0.182968</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>0.0436753</td>
<td>0.208986</td>
</tr>
</tbody>
</table>

The latency to the preceding has a large effect size with a 400 ms difference between the smallest and largest predictor values, the corresponding difference for the frequency effect was only 50 ms.

The contrast coefficient for \(\text{Conditionheid}\) changes sign when accuracy and response latency to the prime itself, 40 trials back in the experiment, are taken into account.

```r
> print(lmer(RT ~ log(RTmin1) + ResponseToPrime + RTtoPrime + Base Frequency + Condition + (1|Word) + (1|Subject)), cor = FALSE)
```

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>(Intercept)</td>
<td>0.00049959</td>
<td>0.022351</td>
</tr>
<tr>
<td>Subject</td>
<td>(Intercept)</td>
<td>0.02400862</td>
<td>0.154928</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>0.03340644</td>
<td>0.182774</td>
</tr>
</tbody>
</table>

```

we observe significant facilitation from long-distance identity priming. For a follow-up experiment using self-paced reading of continuous text, latencies were likewise codetermined by the reading latencies to the words preceding in the discourse, as well as by the reading latency for the prime. Traditional averaging procedures applied to these data would either report a null effect (for self-paced reading) or would lead to a completely wrong interpretation of the data (lexical decision). Mixed-effects modeling allows us to avoid these pitfalls, and makes it possible to obtain substantially improved insight into the structure of one’s experimental data.

### Some common designs

Having illustrated the important analytical advantages offered by mixed-effects modeling with crossed random effects for subjects and items, we now turn to consider how mixed-effects modeling compares to traditional analysis of variance and random regression. Raaijmakers, Schrijnemakers, and Gremmen (1999) discuss two common factorial experimental designs and their analyses. In this section, we first report simulation studies using their designs, and compare the performance of current standards with the performance of mixed-effects models. Simulations were run in R (version 2.4.0) (R development core team, 2007) using the lme4 package of Bates and Sarkar (2007) (see also Bates,
The code for the simulations is available in the languageR package in the CRAN archives (http://cran.r-project.org, see Baayen, 2008). We then illustrate the robustness of mixed-effects modeling to missing data for a split-plot design, and then pit mixed-effects regression against random regression, as proposed by Lorch and Myers (1990).

**A design traditionally requiring quasi-F ratios**

A constructed dataset discussed by Raaijmakers et al. (1999) comprises 64 observations with 8 subjects and 8 items. Items are nested under treatment: 4 items are presented with a short SOA, and 4 with a long SOA. Subjects are crossed with item. A quasi-F test, the test recommended by Raaijmakers et al. (1999), based on the mean squares in the mean squares decomposition shown in Table 3 shows that the effect of SOA is not significant ($F(1,1025,9.346) = 1.702, p = 0.224$). It is noteworthy that the model fits 64 data points with the help of 72 parameters, 6 of which are inestimable.

The present data set is available in the languageR package as quasif. We fit a mixed effects model to the data with

```
> quasif = lmer(RT ~ SOA + (1|Item) + (1 + SOA|Subject), data = quasif)
```

and inspect the estimated parameters with

```
> summary(quasif)
Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>(Intercept)</td>
<td>448.29</td>
<td>21.73</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>(Intercept)</td>
<td>861.99</td>
<td>29.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOAshort</td>
<td>502.65</td>
<td>22.42</td>
<td>-0.813</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>100.31</td>
<td>10.016</td>
<td></td>
</tr>
</tbody>
</table>

number of obs: 64, groups: Item, 8; Subject, 8

Fixed effects:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std.Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>540.91</td>
<td>14.93</td>
<td>36.23</td>
</tr>
<tr>
<td>SOAshort</td>
<td>22.41</td>
<td>22.38</td>
<td>1.31</td>
</tr>
</tbody>
</table>

The small t-value for the contrast coefficient for SOA shows that this predictor is not significant. This is clear as well from the summary of the fixed effects produced by pvals.fnc (available in the languageR package), which lists the estimates, their MCMC means, the corresponding HPD intervals, the two-tailed MCMC probability, and the two-tailed probability derived from the t-test using, as mentioned above, the upper bound for the degrees of freedom.

```
> pvals.fnc(quasif, nsim = 10000)
```

|                         | Estimate | MCMCmean | HPD95lower | HPD95upper | pMCMC  | Pr(>|t|) |
|-------------------------|----------|----------|------------|------------|--------|---------|
| (Intercept)             | 540.91   | 540.85   | 498.58     | 583.50     | 0.0001 | 0.0000  |
| SOAshort                | 22.41    | 22.38    | -32.88     | 76.29      | 0.3638 | 0.1956  |

The model summary lists four random effects: random intercepts for participants and for items, by-participant random slopes for SOA, and the residual error. Each random effect is paired with an estimate of the standard deviation that characterizes the spread of the random effects for the slopes and intercepts. Because the by-participant BLUPs for slopes and intercepts are paired observations, the model specification that we used here allows for these two random variables to be correlated. The estimate of this correlation ($r = -0.813$) is the final parameter of the present mixed effects model.

The p-value for the t-test obtained with the mixed-effects model is slightly smaller than that produced by the quasi-F test. However, for the present small data set the MCMC p-value is to be used, as the p-value with the above mentioned upper bound for the degrees of freedom is anticonservative. To see this, consider Table 4, which summarizes Type I error rate and power across simulated data sets, 1000 with and 1000 without an effect of SOA. The number of simulation runs is kept small on purpose: These simulations are provided to illustrate only main trends in power and error rate.

For each simulated data set, five analyses were conducted: a mixed-effects analysis with the anticonservative p-value based on the t-test and the appropriate p-value based on 10,000 MCMC samples generated from the posterior distribution of the parameters of the fitted mixed-effects model, a quasi-F test, a by-participant analysis, a by-item analysis, and an analysis that accepted the effect of SOA to be significant only if both the F1 and the F2 test were significant ($F_1 + F_2$, compare Forster & Dickinson, 1976). This anticonservatism of the t-test is clearly visible in Table 4.

The only procedures with nominal Type I error rates are the quasi-F test and the mixed-effects model with MCMC sampling. For data sets with few observations, the quasi-F test emerges as a good choice with somewhat greater power.
Table 4
Proportions (for 1000 simulation runs) of significant treatment effects for mixed-effects models (lmer), quasi-F tests, by-participant and by-item analyses, and the combined F1 and F2 test, for simulated models with and without a treatment effect for a data set with 8 subjects and 8 items

<table>
<thead>
<tr>
<th></th>
<th>lmer: $p(t)$</th>
<th>lmer: $p(MCMC)$</th>
<th>quasi-F</th>
<th>By-subject</th>
<th>By-item</th>
<th>F1 + F2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Without treatment effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$z = 0.05$</td>
<td>0.088</td>
<td>0.032</td>
<td>0.055</td>
<td>0.310</td>
<td>0.081</td>
<td>0.079</td>
</tr>
<tr>
<td>$z = 0.01$</td>
<td>0.031</td>
<td>0.000</td>
<td>0.005</td>
<td>0.158</td>
<td>0.014</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>With treatment effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$z = 0.05$</td>
<td>0.16</td>
<td>0.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$z = 0.01$</td>
<td>0.04</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Markov Chain Monte Carlo estimates of significance are denoted by $MCMC$. Power is tabulated only for models with nominal Type 1 error rates. Too high Type 1 error rates are shown in bold.

Table 5
Proportions (for 1000 simulation runs) of significant treatment effects for mixed-effects models (lmer), quasi-F tests, by-participant and by-item analyses, and the combined F1 and F2 test, for simulated models with and without a treatment effect for 20 subjects and 40 items

<table>
<thead>
<tr>
<th></th>
<th>lmer: $p(t)$</th>
<th>lmer: $p(MCMC)$</th>
<th>quasi-F</th>
<th>Subject</th>
<th>Item</th>
<th>F1 + F2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Without treatment effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$z = 0.05$</td>
<td>0.055</td>
<td>0.027</td>
<td>0.052</td>
<td>0.238</td>
<td>0.102</td>
<td>0.099</td>
</tr>
<tr>
<td>$z = 0.01$</td>
<td>0.013</td>
<td>0.001</td>
<td>0.009</td>
<td>0.120</td>
<td>0.036</td>
<td>0.036</td>
</tr>
<tr>
<td><strong>With treatment effect</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$z = 0.05$</td>
<td>0.823</td>
<td>0.681</td>
<td>0.809</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$z = 0.01$</td>
<td>0.618</td>
<td>0.392</td>
<td>0.587</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Power is tabulated only for models with nominal Type 1 error rates. Too high Type 1 error rates are shown in bold.

Most psycholinguistic experiments yield much larger numbers of data points than in the present example. Table 5 summarizes a second series of simulations in which we increased the number of subjects to 20 and the number of items to 40. As expected, the Type I error rate for the mixed-effects models evaluated with tests based on $p$-values using the $t$-test are now in accordance with the nominal levels, and power is perhaps slightly larger than the power of the quasi-F test. Evaluation using MCMC sampling is conservative for this specific fully balanced example. Depending on the costs of a Type I error, the greater power of the $t$-test may offset its slight anti-conservatism. In our experience, the difference between the two $p$-values becomes very small for data sets with thousands instead of hundreds of observations. In analyses where MCMC-based evaluation and $t$-based evaluation yield a very similar verdict across coefficients, exceptional disagreement, with MCMC sampling suggesting clear non-significance and the $t$-test suggesting significance, is a diagnostic of an unstable and suspect parameter. This is often confirmed by inspection of the parameter’s posterior density.

It should be kept in mind that real life experiments are characterized by missing data. Whereas the quasi-F test is known to be vulnerable to missing data, mixed-effects models are robust in this respect. For instance, in 1000 simulation runs (without an effect of SOA) in which 20% of the datapoints are randomly deleted before the analyses are performed, the quasi-F test emerges as slightly conservative (Type I error rate: 0.045 for $z = 0.05, 0.006$ for $z = 0.010$), whereas the mixed-effects model using the $t$ test is on target (Type I error rate: 0.052 for $z = 0.05, 0.010$ for $z = 0.01$). Power is slightly greater for the mixed analysis evaluating probability using the $t$-test ($z = 0.05: 0.84$ versus $0.81$ for the quasi-F test; $z = 0.01: 0.57$ versus $0.54$). See also, e.g., Pinheiro and Bates (2000).

A Latin Square design

Another design discussed by Raaijmakers and colleagues is the Latin Square. They discuss a second constructed data set, with 12 words divided over 3 lists with 4 words each. These lists were rotated over participants, such that a given participant was exposed to a list for only one of three SOA conditions. There were 3 groups of 4 participants, each group of participants was exposed to unique combinations of list and SOA. Raaijmakers and colleagues recommend a by-participant analysis that proceeds on the basis of means obtained by averaging over the words in the lists. An analysis of variance is performed on the resulting data set which lists, for each participant, three means; one for each SOA condition. This gives rise to the ANOVA decomposi-
tion shown in Table 6. The $F$ test compares the mean squares for SOA with the mean squares of the interaction of SOA by List, and indicates that the effect of SOA is not statistically significant ($F(2, 2) = 1.15$, $p = 0.465$). As the interaction of SOA by List is not significant, Raaijmakers et al. (1999) pool the interaction with the residual error. This results in a pooled error term with 20 degrees of freedom, an $F$-value of 0.896, and a slightly reduced $p$-value of 0.42.

A mixed-effects analysis of the same data set (available as latinsquare in the languageR package) obviates the need for prior averaging. We fit a sequence of models, decreasing the complexity of the random effects structure step by step.

The likelihood ratio tests show that the model with Subject and Word as random effects has the right level of complexity for this data set. SOA compared to the reference level, long SOA. Inspection of the corresponding $p$-values shows that the $p$-value based on the $t$-test and that based on MCMC sampling are very similar, and the same holds for the $p$-value produced by the $F$-test for the factor SOA.

The summary of this model lists the three random effects and the corresponding parameters: the variances (and standard deviations) for the random intercepts for subjects and items, and for the residual error. The fixed-effects part of the model provides estimates for the intercept and for the contrasts for medium and short SOA compared to the reference level, long SOA.

The likelihood ratio tests show that the model with Subject and Word as random effects has the right level of complexity for this data set.
compared to the F1 analysis proposed by Raaijmakers et al. (1999), as illustrated in Table 7, which lists Type I error rate and power for 1000 simulation runs without and with an effect of SOA. Simulated datasets were constructed using the parameters given by Latin-square.lmer4. The upper half of Table 7 shows power and Type I error rate for the situation in which the F1 analysis includes the interaction of SOA by List, the lower half reports the case in which this interaction is pooled with the residual error. Even for the most powerful test suggested by Raaijmakers et al. (1999), the mixed-effects analysis emerges with slightly better power, while maintaining the nominal Type-I error rate. Further pooling of non-explanatory parameters in the F1 approach may be expected to lead to further convergence of power. The key point that we emphasize here is that the mixed-effects approach obtains this power without prior averaging. As a consequence, it is only the mixed-effects approach that affords the possibility of bringing predictors for longitudinal effects and inter-trial dependencies into the model. Likewise, the possibility of bringing covariates gauging properties of the individual words into the model is restricted to the mixed-effects analysis.

A split-plot design

Another design often encountered in psycholinguistic studies is the split plot design. Priming studies often make use of a counterbalancing procedure with two sets of materials. Words are primed by a related prime in List A and by an unrelated prime in List B, and vice versa. Different subjects are tested on each list. This is a split-plot design, in the sense that the factor List is between subjects and the factor Priming within subjects. The following example presents an analysis of an artificial dataset (dat, available as splitplot in the languageR package) with 20 subjects, 40 items. A series of likelihood ratio tests on a sequence of models with decreasing complex random effects structure shows that a model with random intercepts for subject and item suffices.

```r
dat.lmer1 = lmer(RT ~ list * priming + (1 + priming | subjects) + (1 + list | items), data = dat) 
dat.lmer2 = lmer(RT ~ list * priming + (1 + priming | subjects) + (1 | items), data = dat) 
dat.lmer3 = lmer(RT ~ list * priming + (1 | subjects) + (1 | items), data = dat) 
dat.lmer4 = lmer(RT ~ list * priming + (1 | subjects), data = dat) 
anova(dat.lmer1, dat.lmer2, dat.lmer3, dat.lmer4)
```

Table 7
Proportions (out of 1000 simulation runs) of significant F-tests for a Latin Square design with mixed-effects models (lmer) and a by-subject analysis (F1)

<table>
<thead>
<tr>
<th></th>
<th>Without SOA</th>
<th>With SOA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lmer: p(F)</td>
<td>lmer: p(MCMC)</td>
</tr>
<tr>
<td>z = 0.05</td>
<td>With</td>
<td>0.055</td>
</tr>
<tr>
<td>z = 0.01</td>
<td>With</td>
<td>0.011</td>
</tr>
<tr>
<td>z = 0.05</td>
<td>Without</td>
<td>0.038</td>
</tr>
<tr>
<td>z = 0.01</td>
<td>Without</td>
<td>0.010</td>
</tr>
</tbody>
</table>

The upper part reports simulations in which the F1 analysis includes the interaction of List by SOA (With), the lower part reports simulations in which for the F1 analysis this interaction is absent (Without).

Table 8
Type I error rate and power for mixed-effects modeling of 1000 simulated data sets with a split-plot design, for the full data set and a data set with 20% missing data

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>z = 0.05</td>
<td>0.046</td>
<td>(0.046)</td>
</tr>
<tr>
<td>z = 0.01</td>
<td>0.013</td>
<td>(0.011)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I error rate</td>
<td>0.999</td>
<td>(0.999)</td>
</tr>
<tr>
<td>Power</td>
<td>0.993</td>
<td>(0.993)</td>
</tr>
</tbody>
</table>

Evaluation based on Markov chain Monte Carlo sampling are listed in parentheses.
The estimates are close to the parameters that generated the simulated data: \( r_i = 20, r_s = 50, r_t = 80, b_{\text{int}} = 400, b_{\text{priming}} = 30, b_{\text{list}} = 18.5, b_{\text{list:priming}} = 0. \)

Table 8 lists power and Type I error rate with respect to the priming effect for 1000 simulation runs with a mixed-effect model, run once with the full data set, and once with 20% of the data points randomly deleted, using the same parameters that generated the above data set. It is clear that with the low level of by-observation noise, the presence of a priming effect is almost always detected. Power decreases only slightly for the case with missing data. Even though power is at ceiling, the Type I error rate is in accordance with the nominal levels. Note the similarity between evaluation of significance based on the (anticonservative) \( t \)-test and evaluation based on Markov chain Monte Carlo sampling. This example illustrates the robustness of mixed effects models with respect to missing data: The present results were obtained without any data pruning and without any form of imputation.

### A multiple regression design

Multiple regression designs with subjects and items, and with predictors that are tied to the items (e.g., frequency and length for items that are words) have traditionally been analyzed in two ways. One approach aggregates over subjects to obtain item means, and then proceeds with standard ordinary least squares regression. We refer to this as by-item regression. Another approach, advocated by Lorch and Myers (1990), is to fit separate regression models to the data sets elicited from the individual participants. The significance of a given predictor is assessed by means of a one-sample \( t \)-test applied to the coefficients of this predictor in the individual regression models. We refer to this procedure as by-participant regression. It is also known under the name of random regression. (From our perspective, these procedures combine precise and imprecise information on an equal footing.) Some studies report both by-item and by-participant regression models (e.g., Alegre & Gordon, 1999).

The by-participant regression is widely regarded as superior to the by-item regression. However, the by-participant regression does not take item-variability into account. To see this, compare an experiment in which each participant responds to the same set of words to an experiment in which each participant responds to a different set of words. When the same lexical predictors are used in both experiments, the by-participant analysis proceeds in exactly the same way for both. But whereas this approach is correct for the second experiment, it ignores a systematic source of variation in the case of the first experiment.

A simulation study illustrates that ignoring item variability that is actually present in the data may lead to unacceptably high Type I error rates. In this simulation study, we considered three predictors, \( X, Y \) and \( Z \) tied to 20 items, each of which was presented to 10 participants. In one set of simulation runs, these predictors had beta weights 2, 6 and 4. In a second set of simulation runs, the beta weight for \( Z \) was set to zero. We were interested in the power and Type I error rates for \( Z \) for by-participant and for by-item regression, and for two different mixed-effects models. The first mixed-effects model that we considered included crossed random effects for par-
Table 9
Proportion of simulation runs (out of 1000) in which the coefficients for the intercept and the predictors X, Y, and Z are reported as significantly different from zero according to four multiple regression models

<table>
<thead>
<tr>
<th>Model</th>
<th>(\beta_Z = 0)</th>
<th>(\beta_Z = 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{lmerS: p(t)})</td>
<td>0.609</td>
<td>0.597</td>
</tr>
<tr>
<td>(\text{subj: p(MMC)})</td>
<td>0.699</td>
<td>0.594</td>
</tr>
<tr>
<td>(\text{item})</td>
<td>0.380</td>
<td>0.925</td>
</tr>
<tr>
<td>(\text{Subject Residual})</td>
<td>0.503</td>
<td>0.485</td>
</tr>
<tr>
<td>(\text{Y})</td>
<td>0.553</td>
<td>0.487</td>
</tr>
<tr>
<td>(\text{Z})</td>
<td>0.982</td>
<td>0.979</td>
</tr>
<tr>
<td>(\text{X})</td>
<td>0.238</td>
<td>0.867</td>
</tr>
<tr>
<td>(\text{Y})</td>
<td>0.239</td>
<td>0.869</td>
</tr>
<tr>
<td>(\text{Z})</td>
<td>0.269</td>
<td>0.868</td>
</tr>
<tr>
<td>(\text{X})</td>
<td>0.013</td>
<td>0.295</td>
</tr>
</tbody>
</table>

The standard deviation for the subject random effect is slightly underestimated.

We note that for real datasets, mixed-effects regression offers the possibility to include not only item-bound predictors, but also predictors tied to the subjects, as well as predictors capturing inter-trial dependencies and longitudinal effects.

Further issues

Some authors, e.g., Quené and Van den Bergh (2004), have argued that in experiments with subjects and items, items should be analyzed as nested under subjects. The nesting of items under participants creates a hierarchical mixed-effects model. Nesting is argued to be justified on the grounds that items may vary in familiarity across participants. For instance, if items are words, than lexical familiarity is known to vary considerably across occupations (see, e.g., Gardner, Rothkopf, Lapan, & Lafferty, 1987). Technically, however, nesting amounts to the strong assumption that there need not be any commonality at all for a given item across participants.

This strong assumption is justified only when the predictors in the regression model are treatments administered to items that otherwise do not vary on dimensions that might in any way affect the outcome of the experiment. For many linguistic items, predictors are intrinsically bound to the items. For instance, when items are words, predictors such as word frequency and word length are not treatments administered to items. Instead, these predictors gauge aspects of a word’s lexical properties. Furthermore, for many current studies it is unlikely that they fully exhaust all properties that co-determine lexical processing. In these circumstances, it is highly likely that there is a non-negligible residue of item-bound properties that are not brought into the model formulation. Hence a random effect for word should be considered seriously. Fortunately, mixed-effects models allow the researcher to explicitly test whether a random effect for Item is required by means of a likelihood ratio test comparing a model with and without a random effect for item. In our experience, such tests almost invariably show that a random effect for item is required, and the resulting models provide a tighter fit to the data.

Mixed-effects regression with crossed random effects for participants and items have further advantages to offer. One advantage is shrinkage estimates for the
BLUPS (the subject and item specific adjustments to intercepts and slopes), which allow enhanced prediction for these items and subjects (see, e.g., Baayen, 2008, for further discussion). Another important advantage is the possibility to include simultaneously predictors that are tied to the items (e.g., frequency, length) and predictors that are tied to participants (e.g., handedness, age, gender). Mixed-effects models have also been extended to generalized linear models and can hence be used efficiently to model binary response data such as accuracy in lexical decision (see Jaeger, this volume).

To conclude, we briefly address the question of the extent to which an effect observed to be significant in a mixed-effects analysis generalizes across both subjects and items (see Forster, this issue). The traditional interpretation of the F1 (by-subject) and F2 (by-item) analyses is that significance in the F1 analysis would indicate that the effect is significant for all subjects, and that the F2 analysis would indicate that the effect holds for all items. We believe this interpretation is incorrect. In fact, even if we replace the F1+F2 procedure by a mixed-effects model, the inference that the effect would generalize across all subjects and items remains incorrect. The fixed-effect coefficients in a mixed-effect model are estimates of the intercept, slopes (for numeric predictors) or contrasts (for factors) in the population for the average, unknown subject and the average, unknown item. Individual subjects and items may have intercepts and slopes that diverge considerably from the population means. For ease of exposition, we distinguish three possible states of affairs for what in the traditional terminology would by described as an Effect by Item interaction.

First, it is conceivable that the BLUPS for a given fixed-effect coefficient, when added to that coefficient, never change its sign. In this situation, the effect indeed generalizes across all subjects (or items) sampled in the experiment. Other things being equal, the partial effect of the predictor quantified by this coefficient will be highly significant.

Second, situations arise in which adding the BLUPS to a fixed coefficient results in a majority of by-subject (or by-item) coefficients that have the same sign as the population estimate, in combination with a relatively small minority of by-subject (or by-item) coefficients with the opposite sign. The partial effect represented by the population coefficient will still be significant, but there will be less reason for surprise. The effect generalizes to a majority, but not to all subjects or items. Nevertheless, we can be confident about the magnitude and sign of the effect on average, for unknown subjects or items, if the subjects and items are representative of the population from which they are sampled.

Third, the by-subject (or by-item) coefficients obtained by taking the BLUPS into account may result in a set of coefficients with roughly equal numbers of coefficients that are positive and coefficients that are negative. In this situation, the main effect (for a numeric predictor or a binary contrast) will not be significant, in contradistinction to the significance of the random effect for the slopes or contrasts at issue. In this situation, there is a real and potentially important effect, but averaged across subjects or items, it cancels out to zero.

In the field of memory and language, experiments that do not yield a significant main effect are generally considered to have failed. However, an experiment resulting in this third state of affairs may constitute a positive step forward for our understanding of language and language processing. Consider, by way of example, a pharmaceutical company developing a new medicine, and suppose this medicine has adverse side effects for some, but highly beneficial effects for other patients—patients for which it is an effective life-saver. The company could decide not to market the medicine because there is no main effect. However, they can actually make substantial profit by bringing it on the market with warnings for adverse side effects and proper distributional controls.

Returning to our own field, we know that no two brains are the same, and that different brains have different developmental histories. Although in the initial stages of research the available technology may only reveal the most robust main effects, the more our research advances, the more likely it will become that we will be able to observe systematic individual differences. Ultimately, we will need to bring these individual differences into our theories. Mixed-effect models have been developed to capture individual differences in a principled way, while at the same time allowing generalizations across populations. Instead of discarding individual differences across subjects and items as an uninteresting and disappointing nuisance, we should embrace them. It is not to the advantage of scientific progress if systematic variation is systematically ignored.

Hierarchical models in developmental and educational psychology

Thus far, we have focussed on designs with crossed random effects for subjects and items. In educational and developmental research, designs with nested random effects are often used, such as the natural hierarchy formed by students nested within a classroom (Goldstein, 1987). Such designs can also be handled by mixed-effects models, which are then often referred to as hierarchical linear models or multilevel models.

Studies in educational settings are often focused on learning over time, and techniques developed for this type of data often attempt to characterize how individuals’ performance or knowledge changes over time, termed the analysis of growth curves (Goldstein, 1987, 1995; Goldstein et al., 1993; Nutall, Goldstein, Prosser,
& Rasbash, 1989; Willet, Singer, & Martin, 1998). Examples of this include the assessment of different teaching techniques on students performance (Aitkin, Anderson, & Hinde, 1981), and the comparison of the effectiveness of different schools (Aitkin & Longford, 1986). Goldstein et al. (1993) used multilevel techniques to study the differences between schools and students when adjusting for pre-existing differences when students entered classes. For a methodological discussion of the use of these models, see the collection of articles in the Summer 1995 special issue of Journal of Educational and Behavioral Statistics on hierarchical linear models, e.g., Kreft (1995). Singer (1998) provides a practical introduction to multilevel models including demonstration code, and Collins (2006) provides a recent overview of issues in longitudinal data analysis involving these models. Finally, Fielding and Goldstein (2006) provide a comprehensive overview of multilevel and cross-classified models applied to education research, including a brief software review. West et al. (2007) provide a comprehensive software review for nested mixed-effects models.

These types of models are also applicable to psycholinguistic research, especially in studies of developmental change. Individual speakers from a language community are often members of a hierarchy, e.g., language:dialect:family:speaker, and many studies focus on learning or language acquisition, and thus analysis of change or development is important. Huttenlocher, Haight, Bryk, and Seltzer (1991) used multilevel models to assess the influence of parental or caregiver speech on vocabulary growth, for example. Boyle and Willms (2001) provide an introduction to the use of multilevel models to study developmental change, with an emphasis on growth curve modeling and discrete outcomes. Raudenbush (2001) reviews techniques for analyzing longitudinal designs in which repeated measures are used. Recently, Misangyi, LePine, Algina, and Goeddeke (2006) compared repeated measures regression to multivariate ANOVA (MANOVA) and multilevel analysis in research designs typical for organizational and behavioral research, and concluded that multilevel analysis can provide equivalent results as MANOVA, and in cases where specific assumptions about variance-covariance structures could be made, or in cases where missing values were present, that multilevel modeling is a better analysis strategy and in some cases a necessary strategy (see also Kreft & de Leeuw, 1998 and Snijders & Bosker, 1999).

Finally, a vast body of work in educational psychology concerns test construction and the selection of test items (Lord & Novick, 1968). Although it is beyond the scope of this article to review this work, it should be noted that work within generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) has been concerned with the problem of crossed subject and item factors using random effects models (Schroeder & Hakstian, 1990). For a recent application of the software considered here to item response theory, see Doran, Bates, Bliise, and Dowling (2007), and for the application of hierarchical models to joint response type and response time measures, see Fox, Entink, and van der Linden (2007).

Mixed-effects models in neuroimaging

In neuroimaging, two-level or mixed effects models are now a standard analysis technique (Friston et al., 2002a, 2002b; Worsley et al., 2002), and are used in conjunction with Gaussian Random Field theory to make inferences about activity patterns in very large data sets (voxels from fMRI scans). These techniques are formally comparable to the techniques that are advocated in this paper (Friston, Stephan, Lund, Morcom, & Kiebel, 2005). Interestingly, however, the treatment of stimuli as random effects has not been widely addressed in the imaging and physiological community, until recently (Bedny, Aguirre, & Thompson-Schill, 2007).

In imaging studies that compare experimental conditions, for example, statistical parameter maps (spm; Friston et al., 1995) are calculated based on successively recorded time series for the different experimental conditions. A hypothesized hemodynamic response function is convolved with a function that encodes the experimental design matrix, and this forms a regressor for each of the time series in each voxel. Significant parameters for the regressors are taken as evidence of activity in the voxels that exhibit greater or less activity than is expected based on the null hypothesis of no activity difference between conditions. The logic behind these tests is that a rejection of the null hypothesis for a region is evidence for a difference in activity in that region.

Neuroimaging designs are often similar to cognitive psychology designs, but the dimension of the response variable is much larger and the nature of the response has different statistical properties. However, this is not crucial for the application of mixed effects models. In fact, it shows the technique can scale to problems that involve very large datasets.

A prototypical case of a fixed effects analysis in fMRI would test whether a image contrast is statistically significant within a single subject over trials. This would be analogous to a psychophysics experiment using only a few participants, or a patient case study. For random effect analysis the parameters calculated from the single participants are used in a mixed model to test whether a contrast is significant over participants, in order to test whether the contrasts reflects a difference in the population from which the participants were sampled. This is analogous to how cognitive psychology experimenters treat mean RTs, for example. A common analysis strategy is to calculate a single parameter for each participant.
in an RT study, and then analyze this data in (what in the neuroimaging community is called) a random effects analysis.

The estimation methods used to calculate the statistical parameters of these models include Maximum Likelihood or Restricted Maximum Likelihood, just as in the application of the multilevel models used in education research described earlier. One reason that these techniques are used is to account for correlation between successive measurements in the imaging time series. These corrections are similar to corrections familiar to psychologists for non-sphericity (Greenhouse & Geisser, 1958).

Similar analysis concerns are present within electro-physiology. In the past, journal policy in psychophysical research has dealt with the problems posed by repeated measures experimental designs by suggesting that researchers adopt statistical procedures that take into account the correlated data obtained from these designs (Jennings, 1987; Vasey & Thayer, 1987). Mixed effects models are less commonly applied in psychophysical research, as the most common techniques are the traditional univariate ANOVA with adjustments or multivariate ANOVA (Dien & Santuzzi, 2004), but some researchers have advocated them to deal with repeated measures data. For example, Bagiella, Sloan, and Heitjan (2000) suggest that mixed effects models have advantages over more traditional techniques for EEG data analysis.

The current practice of psychophysicologists and neuroimaging researchers typically ignores the issue of whether linguistic materials should be modeled with fixed or random effect models. Thus, while there are techniques available for modeling stimuli as random effects, it is not yet current practice in neuroimaging and psychophysiology to do so. This represents a tremendous opportunity for methodological development in language-related imaging experiments, as psychologists have considerable experience in modeling stimulus characteristics.

Cognitive psychologists and neuroscientists might reasonably assume that the language-as-a-fixed-effect debate is only a concern when linguistic materials are used, given that most discussion to date has taken place in the context of linguistically-motivated experiments. This assumption is too narrow, however, because naturalistic stimuli from many domains are drawn from populations.

Consider a researcher interested in the electrophysiology of face perception. She designs an experiment to test whether an ERP component such as the N170 in response to faces has a different amplitude in one of two face conditions, normal and scrambled form. She obtains a set of images from a database, arranges them according to her experimental design, and proceeds to present each picture in a face-detection EEG experiment, analogous to the way that a psycholinguist would present words and non-words to a participant in a lexical decision experiment. The images presented in this experiment would be a sample of all possible human faces. It is not controversial that human participants are to be modeled as a random variable in psychological experiments. Pictures of human faces are images of a random variable, presented as stimuli. Thus, it should be no source of controversy that naturalistic face stimuli are also a random variable, and should be modeled as a random effect, just like participants. For the sake of consistency, if human participants, faces, and speech are to be considered random variables, then objects, artifacts, and scenes might just as well be considered random variables (also pointed out by Raaijmakers, 2003).

Any naturalistic stimulus which is a member of a population of stimuli which has not been exhaustively sampled should be considered a random variable for the purposes of an experiment. Note that random in this sense means stochastic, a variable subject to probabilistic variation, rather than randomly sampled. A random sample is one method to draw samples from a population and assign them to experimental condition. However, stimuli may have stochastic characteristics whether or not they are randomly sampled or not. Participants have stochastic characteristics, as well, whether they are randomly sampled or not. Therefore, the present debate about the best way to model random effects of stimuli is wider than previously has been appreciated, and should be seen as part of the debate over the use of naturalistic stimuli in sensory neurophysiology as well (Felsen & Yang, 2005; Ruse & Movshon, 2005).

Concluding remarks

We have described the advantages that mixed-effects models with crossed random effects for subject and item offer to the analysis of experimental data.

The most important advantage of mixed-effects models is that they allow the researcher to simultaneously consider all factors that potentially contribute to the understanding of the structure of the data. These factors comprise not only standard fixed-effects factors typically manipulated in psycholinguistic experiments, but also covariates bound to the items (e.g., frequency, complexity) and the subjects (e.g., age, sex). Furthermore, local dependencies between the successive trials in an experiment can be brought into the model, and the effects of prior exposure to related or identical stimuli (as in long-distance priming) can be taken into account as well. (For applications in eye-movement research, see Klögl, 2007, and Klögl, Risse, & Laubrock, 2007). Mixed-effects models may offer substantially enhanced insight into how subjects are performing in the course of an experiment, for instance, whether they are adjust-
ing their behavior as the experiment proceeds to optimize performance. Procedures requiring prior averaging across subjects or items, or procedures that are limited to strictly factorial designs, cannot provide the researcher with the analytical depth typically provided by a mixed-effects analysis.

For data with not too small numbers of observations, mixed-effects models may provide modest enhanced power, as illustrated for a Latin Square design in the present study. For regression and analysis of covariance, mixed-effects modeling protects against inflated significance for data sets with significant by-item random effects structure. Other advantages of mixed-effects modeling that we have mentioned only in passing are the principled way in which non-independence (asphericity) is handled through the variance-covariance structure of the model, and the provision of shrinkage estimates for the by-subject and by-item adjustments to intercept and slopes, which allows enhanced precision in prediction.

An important property of mixed-effects modeling is that it is possible to fit models to large, unbalanced data sets. This allows researchers to investigate not only data elicited under controlled experimental conditions, but to also study naturalistic data, such as corpora of eye-movement data. Markov chain Monte Carlo sampling from the posterior distribution of the parameters is an efficient technique to evaluate fitted models with respect to the stability of their parameters and to distinguish between robust parameters (with narrow highest posterior density intervals) from superfluous parameters (with very broad density intervals).

Mixed-effects modeling is a highly active research field. Well-established algorithms and techniques for parameter estimation are now widely available. One question that is still hotly debated is the appropriate number of degrees of freedom for the fixed-effects factors. Different software packages make use of or even offer different choices. We have emphasized the importance of Markov chain Monte Carlo sampling as fast and efficient way (compared to, e.g., the bootstrap) to evaluate a model’s parameters. In our experience, p-values based on MCMC sampling and p-values based on the upper bound of the degrees of freedom tend to be very similar for all but the smallest samples.

An important goal driving the development of the lme4 package in R, the software that we have introduced and advocated here, is to make it possible to deal realistically with the parameters of models fit to large, unbalanced data sets. Bates (2007a) provides an example of a data set with about 1.7 million observations, 55000 “subjects” (distinct students at a major university over a 5 year period) and 7900 “items” (instructors). The data are unbalanced and the subject and item factors are partially crossed. Fitting a simple model with random effects for subject and for item took only about an hour on a fast server computer with substantial memory. Thanks to the possibility of handling very large data sets, we anticipate mixed-effects modeling to become increasingly important for improved modeling of spatial and temporal dependencies in neuroimaging studies, as well as for the study of naturalistic corpus-based data in chronometric tasks and eye-movement research. In short, mixed-effects modeling is emerging not only as a useful but also as an actually useable tool for coming to a comprehensive understanding of the quantitative structure of highly complex data sets.

A note on parameter estimation

The mathematical details of model fitting with mixed effects models are beyond the scope of the present paper (see Bates, 2007, for an introduction), we note here that fitting the model involves finding the right balance between the complexity of the model and faithfulness to the data. Model complexity is determined primarily by the parameters that we invest in the random effects structure, basically the parameters that define the relative variance-covariance matrix \( \Sigma \) in Eq. (10). Interestingly, the profiled deviance function, which is negative twice the log-likelihood of model (10) evaluated at \( \sum, \beta \) and \( \sigma^2 \) for a given set of parameters, can be estimated without having to solve for \( \beta \) or \( b \). The profiled deviance function has two components, one that measures model complexity and one that measures fidelity of the fitted values to the observed data. This is illustrated in Fig. 2.

Each panel has the relative standard deviation of the item random effect (i.e., \( \sigma_i/\sigma \)) on the horizontal axis, and the relative standard deviation of the subject random effect (\( \sigma_s/\sigma \)) on the vertical axis. First consider the rightmost panel. As we allow these two relative standard deviations to increase, the fidelity to the data increases and the deviance (the logarithm of the penalized residual sum of squares) decreases. In the contour plot, darker shades of grey represent greater fidelity and decreased deviance, and it is easy to see that a better fit is obtained for higher values for the item and subject relative standard deviations. However, increasing these relative standard deviations leads to a model that is more complex.2 This is shown in the middle panel, which plots the contours of the model complexity, the logarithm of the determinant of a matrix derived from the random effects matrix \( Z \). Darker shades of grey are now found in the lower left corner, instead of in the upper right corner. The left panel of Fig. 2 shows the compromise between model complexity and fidelity to the data in the form of the deviance function that is minimized at the maxi-

\footnote{The relation between model complexity and the magnitudes of the item and subject relative standard deviations is most easily appreciated by considering the limiting case in which both relative standard deviations are zero. These two parameters can now be removed from the symbolic specification of the model. This reduction in the number of parameters is the familiar index of model simplification.}
Fig. 2. Contours of the profiled deviance as a function of the relative standard deviations of the item random effects (x-axis) and the subject random effects (y-axis). The leftmost panel shows the deviance, the function that is minimized at the maximum likelihood estimates, the middle panel shows the component of the deviance that measures model complexity and the rightmost panel shows the component of the deviance that measures fidelity of the fitted values to the observed data.

References


