Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 59 (2008) 390–412
Journal of Memory and Language
www.elsevier.com/locate/jml

Mixed-effects modeling with crossed random effects for subjects and items

R.H. Baayen [a,*], D.J. Davidson [b], D.M. Bates [c]

[a] University of Alberta, Edmonton, Department of Linguistics, Canada T6G 2E5
[b] Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands
[c] University of Wisconsin, Madison, Department of Statistics, WI 53706-168, USA

Received 15 February 2007; revision received 13 December 2007
Available online 3 March 2008

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.

* Corresponding author. Fax: +1 780 4920806.
E-mail addresses: baayen@ualberta.ca (R.H. Baayen), Doug.Davidson@fcdonders.ru.nl (D.J. Davidson), bates@stat.wisc.edu (D.M. Bates).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.12.005

Clark (1973) illuminated this issue, sparked by the work of Coleman (1964), by showing how language researchers might generalize their results to the larger population of linguistic materials from which they sample by testing for statistical significance of experimental contrasts with participants and items analyses. Clark's oft-cited paper presented a technical solution to this modeling problem, based on statistical theory and computational methods available at the time (e.g., Winer, 1971). This solution involved computing a quasi-F statistic which, in the simplest-to-use form, could be approximated by the use of a combined minimum-F statistic derived from separate participants (F1) and items (F2) analyses. In the *30*+ years since, statistical techniques have expanded the space of possible solutions to this problem, but these techniques have not yet been applied widely in the field of language and memory studies. The present paper discusses an alternative known as a mixed effects model approach, based on maximum likelihood methods that are now in common use in many areas of science, medicine, and engineering (see, e.g., Faraway, 2006; Fielding & Goldstein, 2006; Gilmour, Thompson, & Cullis, 1995; Goldstein, 1995; Pinheiro & Bates, 2000; Snijders & Bosker, 1999).

Software for mixed-effects models is now widely available, in specialized commercial packages such as MLwiN (MLwiN, 2007) and ASReml (Gilmour, Gogel, Cullis, Welham, & Thompson, 2002), in general commercial packages such as SAS and SPSS (the'mixed' procedures), and in the open source statistical programming environment R (Bates, 2007). West, Welch, and Gałlechki (2007) provide a guide to mixed models for five different software packages.

In this paper, we introduce a relatively recent development in computational statistics, namely, the possibility to include subjects and items as crossed, independent, random effects, as opposed to hierarchical or multilevel models in which random effects are assumed to be nested. This distinction is sometimes absent in general treatments of these models, which tend to focus on nested models. The recent textbook by West et al. (2007), for instance, does not discuss models with crossed random effects, although it clearly distinguishes between nested and crossed random effects, and advises the reader to make use of the `lmer()` function in R, the software (developed by the third author) that we introduce in the present study, for the analysis of crossed data.

Traditional approaches to random effects modeling suffer multiple drawbacks which can be eliminated by adopting mixed effect linear models. These drawbacks include (a) deficiencies in statistical power related to the problems posed by repeated observations, (b) the lack of a flexible method of dealing with missing data, (c) disparate methods for treating continuous and categorical responses, as well as (d) unprincipled methods 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 no by-item random slopes) for SOA.

Formally, this dataset is summarized in (1).

$$\mathbf{y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{S}_i\mathbf{s}_i + \mathbf{W}_j\mathbf{w}_j + \boldsymbol{\epsilon}_{ij} \quad (1)$$

The vector $\mathbf{y}_{ij}$ represents the responses of subject i to item j. In the present example, each of the vectors $\mathbf{y}_{ij}$ 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

Subj	Item	SOA	RT	Fixed		ItemInt	Random		Res
				Int	SOA		SubInt	SubSOA	
s1	w1	Long	466	522.2	0	−28.3	−26.2	0	−2.0
s1	w2	Long	520	522.2	0	14.2	−26.2	0	9.8
s1	w3	Long	502	522.2	0	14.1	−26.2	0	−8.2
s1	w1	Short	475	522.2	−19	−28.3	−26.2	11	15.4
s1	w2	Short	494	522.2	−19	14.2	−26.2	11	−8.4
s1	w3	Short	490	522.2	−19	14.1	−26.2	11	−11.9
s2	w1	Long	516	522.2	0	−28.3	29.7	0	−7.4
s2	w2	Long	566	522.2	0	14.2	29.7	0	0.1
s2	w3	Long	577	522.2	0	14.1	29.7	0	11.5
s2	w1	Short	491	522.2	−19	−28.3	29.7	−12.5	−1.5
s2	w2	Short	544	522.2	−19	14.2	29.7	−12.5	8.9
s2	w3	Short	526	522.2	−19	14.1	29.7	−12.5	−8.2
s3	w1	Long	484	522.2	0	−28.3	−3.5	0	−6.3
s3	w2	Long	529	522.2	0	14.2	−3.5	0	−3.5
s3	w3	Long	539	522.2	0	14.1	−3.5	0	6.0
s3	w1	Short	470	522.2	−19	−28.3	−3.5	1.5	−2.9
s3	w2	Short	511	522.2	−19	14.2	−3.5	1.5	−4.6
s3	w3	Short	528	522.2	−19	14.1	−3.5	1.5	13.2

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.

$$\mathbf{X}_{ij} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \tag{2}$$

and is the same for all subjects i and items j. The design matrix is multiplied by the vector of population coefficients $\boldsymbol{\beta}$. Here, this vector takes the form

$$\boldsymbol{\beta} = \begin{pmatrix} 522.2 \\ -19.0 \end{pmatrix} \tag{3}$$

where 522.2 is the coefficient for the intercept, and -19 the contrast for the short as opposed to the long SOA. The result of this multiplication is a vector that again is identical for each combination of subject and item:

$$\mathbf{X}_{ij}\boldsymbol{\beta} = \begin{pmatrix} 522.2 \\ 503.2 \end{pmatrix} \tag{4}$$

It provides the group means for the long and short SOA. These group means constitute the model's best guess about the expected latencies for the population, i.e., for unseen subjects and unseen items.

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 $\mathbf{S}_i$ matrix (in this example) is a full copy of the $\mathbf{X}_{ij}$ 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,

$$\mathbf{S}_{1j}\mathbf{s}_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} -26.2 \\ 11.0 \end{pmatrix} = \begin{pmatrix} -26.2 \\ -15.2 \end{pmatrix}, \tag{5}$$

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 $\mathbf{W}_j$ matrix is again a copy of the design matrix $\mathbf{X}_{ij}$. 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 $\mathbf{w}_j$ 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.

$$\mathbf{W}_j\mathbf{w}_1 = \begin{pmatrix}1\\1\end{pmatrix}(-28.3) = \begin{pmatrix}-28.3\\-28.3\end{pmatrix} \tag{6}$$

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

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

$$\mathbf{y}_1 = \mathbf{y}_{1j} = \mathbf{X}_{1j}\boldsymbol{\beta} + \mathbf{S}\mathbf{s}_1 + \mathbf{W}\mathbf{w}_j + \boldsymbol{\epsilon}_{1j} = \begin{pmatrix}(522.2 + 0) + (-26.2 + 0) + (-28.3) + (-2.0)\\(522.2 + 0) + (-26.2 + 0) + (14.2) + (9.8)\\(522.2 + 0) + (-26.2 + 0) + (14.1) + (-8.2)\\(522.2 - 19) + (-26.2 + 11) + (-28.3) + (15.4)\\(522.2 - 19) + (-26.2 + 11) + (14.2) + (-8.4)\\(522.2 - 19) + (-26.2 + 11) + (14.1) + (11.9)\end{pmatrix} \tag{7}$$

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

$$\mathbf{y}_1 = \begin{pmatrix}(522.2 - 26.2 - 28.3) + (0 + 0) + (-2.0)\\(522.2 - 26.2 + 14.2) + (0 + 0) + (9.8)\\(522.2 - 26.2 + 14.1) + (0 + 0) + (-8.2)\\(522.2 - 26.2 - 28.3) + (-19 + 11) + (15.4)\\(522.2 - 26.2 + 14.2) + (-19 + 11) + (-8.4)\\(522.2 - 26.2 + 14.1) + (-19 + 11) + (11.9)\end{pmatrix} \tag{8}$$

In this equation for $\mathbf{y}_1$ the presence of by-subject random slopes for SOA and the absence of by-item random slopes for SOA is clearly visible.

The subject matrix $\mathbf{S}$ and the item matrix $\mathbf{W}$ can be combined into a single matrix often written as $\mathbf{Z}$, and the subject and item random effects $\mathbf{s}$ and $\mathbf{w}$ can likewise be combined into a single vector generally referenced as $\mathbf{b}$, leading to the general formulation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}. \tag{9}$$

To complete the model specification, we need to be precise about the random effects structure of our data. Recall that a random variable is defined as a normal variate with zero mean and unknown standard deviation. Sample estimates (derived straightforwardly from Table 1) for the standard deviations of the four random effects in our example are $\hat{\sigma}_{s_{\text{int}}} = 28.11$ for the by-subject adjustments to the intercepts, $\hat{\sigma}_{s_{\text{soa}}} = 9.65$ for the by-subject adjustments to the contrast coefficient for SOA, $\hat{\sigma}_i = 24.50$ for the by-item adjustments to the intercept, and $\hat{\sigma}_\epsilon = 8.55$ for the residual error. Because the random slopes and intercept are pairwise tied to the same observational units, they may be correlated. For our data, $\hat{\rho}_{s_{\text{int, soa}}} = -0.71$. 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,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),\ \mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}),\ \mathbf{b} \perp \boldsymbol{\epsilon}, \tag{10}$$

where $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$ is the relative variance-covariance matrix of the random effects in the sense that it is the variance of $\mathbf{b}$ relative to σ^2, the scalar variance of the per-observation noise term $\boldsymbol{\epsilon}$. 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 σ_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):

	Subj	Item	SOA	RT
1	s1	w1	short	475
2	s1	w2	short	494
3	s1	w3	short	490
4	s1	w1	long	466
5	s1	w2	long	520
6	s1	w3	long	502

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

```
> 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

```
> 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 $\hat{\rho}_{s_{\text{int, soa}}}$ of the by-subject random effects for slope and intercept.

A summary of the model is obtained with

```
> 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	
Subj	(Intercept)	803.07	28.338	
	SOAshort	136.46	11.682	−1.000
Residual		102.28	10.113	
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.

Table 2
Comparison of sample estimates and model estimates for the data of Table 1

Parameter	Sample	Model
$\hat{\sigma}_i$	24.50	24.774
$\hat{\sigma}_{s_{\text{int}}}$	28.11	28.338
$\hat{\sigma}_{s_{\text{soa}}}$	9.65	11.681
$\hat{\sigma}_\epsilon$	8.55	10.113
$\hat{\rho}_{s_{\text{int, soa}}}$	−0.71	−1.00

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:

```
> priming.lmer1 = lmer(RT ~ SOA + (1|Item) +
 (1|Subj) + (1| SOA:Subj), data =
 priming)
> 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:

```
> 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 (L_g) should be greater than the likelihood of the more specific model (L_s), and hence the likelihood ratio test statistic $2\log(L_g/L_s) > 0$. 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:

```
> 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	2.9555	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 it clear that the removal of the parameter for the correlation together with the parameter for by-subject random slopes for SOA is justified ($X^2_{(2)} = 2.96, p = 0.228$). The summary of the simplified model

```
> 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		137.35	11.720

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

	Estimate	Std. Error	t value
(Intercept)	522.111	19.602	26.636
SOAshort	−18.889	5.525	−3.419

lists only random intercepts for subject and item, as desired.

The reader may have noted that summaries for model objects fitted with `lmer` 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 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, σ^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 $\hat{\mathbf{b}}$ and the fixed effects $\hat{\boldsymbol{\beta}}$.

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 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 `mcmcsamp` function produces such a sample, for which we plot the estimated densities on a log scale.

```
> 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-

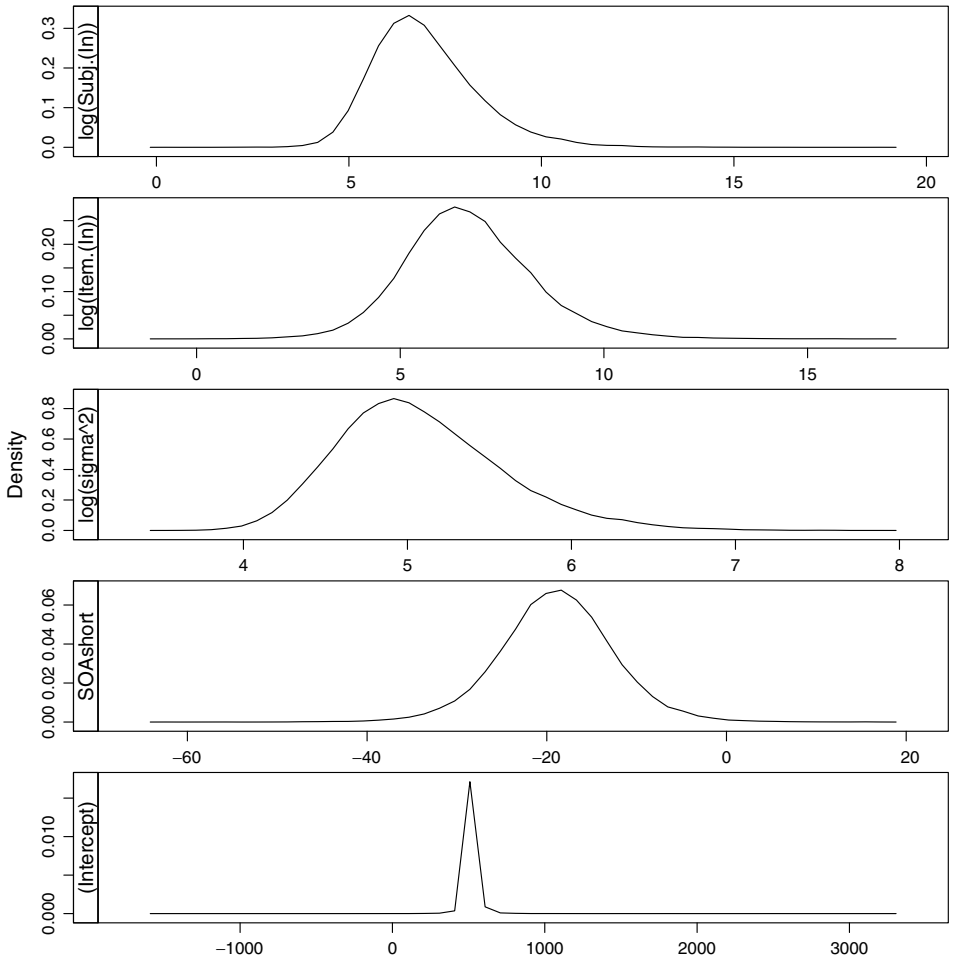

Fig. 1. Empirical density estimates for the Markov chain Monte Carlo sample for the posterior distribution of the parameters in the model for the priming data with random intercepts only (`priming.lmer2`). From top to bottom: $\log \sigma^2_{s_{\text{int}}}$, $\log \sigma^2_{i_{\text{int}}}$ $\log \sigma^2$, β_{soa}, β_{int}.

tribution, 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:

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

	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.[1]

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 χ^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.

[1] For data sets characteristic for studies of memory and language, which typically comprise many hundreds or thousands of observations, the particular value of the number of degrees of freedom is not much of an issue. Whereas the difference between 12 and 15 degrees of freedom may have important consequences for the evaluation of significance associated with a *t* statistic obtained for a small data set, the difference between 612 and 615 degrees of freedom has no noticeable consequences. For such large numbers of degrees of freedom, the *t* distribution has converged, for all practical purposes, to the standard normal distribution. For large data sets, significance at the 5% level in a two-tailed test for the fixed effects coefficients can therefore be gauged informally by checking the summary for whether the absolute value of the *t*-statistic exceeds 2.

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 $\hat{\beta}_i \pm t(\alpha/2, v)s_{\beta_i}$, with the upper bound for the degrees of freedom $(18 - 2 = 16)$ are narrower:

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

	[,1]	[,2]
(Intercept)	479.90683	564.315397
SOAshort	−33.77293	−4.004845

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:

```
> mcmc$random
```

	MCMCmean	HPD95lower	HPD95upper
sigma	12.76	7.947	21.55
Item.(In)	27.54	6.379	140.96
Subj.(In)	32.62	9.820	133.47

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 $H_A : \sigma^2 > 0$ 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 χ^2 does not apply because the parameter value being tested is on the boundary. Therefore, the p-value computed using the χ^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 `primingHeid`PrevRT. After attaching this data set we fit an initial model with `Subject` and `Word` as random effects and priming `Condition` as fixed-effect factor.

```
> 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	0.03127	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 (RTminl) is brought into the model,

```
> print(lmer(RT
log(RTminl) + Condition + (1|Word) + (1|Subject)),
corr = FALSE)
Random effects:
```

Groups	Name	Variance	Std.Dev.
Word	(Intercept)	0.0034623	0.058841
Subject	(Intercept)	0.0334773	0.182968
Residual		0.0436753	0.208986

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

	Estimate	Std. Error	t value
(Intercept)	5.80465	0.22298	26.032
log(RTminl)	0.12125	0.03337	3.633
Conditionheid	0.02785	0.01463	1.903

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 Condition changes sign when accuracy and response latency to the prime itself, 40 trials back in the experiment, are taken into account.

```
> print(lmer(RT log(RTminl) + ResponseToPrime *
RTtoPrime + Condition + (1| Word) +
(1|Subject)), + corr = FALSE)
Random effects:
```

Groups	Name	Variance	Std.Dev.
Word	(Intercept)	0.0013963	0.037367
Subject	(Intercept)	0.0235948	0.153606
Residual		0.0422885	0.205642

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

	Estimate	Std. Error	t value
(Intercept)	4.32436	0.31520	13.720
log(RTminl)	0.11834	0.03251	3.640
ResponseTo Primeincorrect	1.45482	0.40525	3.590
RTtoPrime	0.22764	0.03594	6.334
Conditionheid	−0.02657	0.01618	−1.642
ResponseToPrime incorrect: RTtoPrime	−0.20250	0.06056	−3.344

The table of coefficients reveals that if the prime had elicited a nonword response and the target a word response, response latencies to the target were slowed by some 100 ms, compared to when the prime elicited a word response. For such trials, the response latency to the prime was not predictive for the target. By contrast, the reaction times to primes that were accepted as words were significantly correlated with the reaction time to the corresponding targets.

After addition of log Base Frequency as covariate and trimming of atypical outliers,

```
> priming.lmer = lmer(RT log(RTminl) + ResponseToPrime *
RTtoPrime + Base
Frequency + Condition + (1| + Word) + (1|Subject))
> print(update(priming.lmer,
subset = abs(scale(resid(priming.lmer))) < 2.5),
cor = FALSE)
Random effects:
```

Groups	Name	Variance	Std.Dev.
Word	(Intercept)	0.00049959	0.022351
Subject	(Intercept)	0.02400262	0.154928
Residual		0.03340644	0.182774

number of obs: 815, groups: Word, 40; Subject, 26
Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	4.388722	0.287621	15.259
log(RTminl)	0.103738	0.029344	3.535
ResponseToPrime incorrect	1.560777	0.358609	4.352
RTtoPrime	0.236411	0.032183	7.346
BaseFrequency	−0.009157	0.003590	−2.551
Conditionheid	−0.038306	0.014497	−2.642
ResponseToPrime incorrect: RTtoPrime	−0.216665	0.053628	−4.040

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,

2005). 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 ($F(1.025, 9.346) = 1.702, p = 0.224$). 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:
```

Groups	Name	Variance	Std.Dev.	Corr
Item	(Intercept)	448.29	21.173	
Subject	(Intercept)	861.99	29.360	
	SOAshort	502.65	22.420	−0.813
Residual		100.31	10.016	

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

	Estimate	Std.Error	t value
(Intercept)	540.91	14.93	36.23
SOAshort	22.41	17.12	1.31

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.

Table 3
Mean squares decomposition for the data exemplifying the use of quasi-F ratios in Raaijmakers et al. (1999)

	Df	Sum Sq	Mean Sq
SOA	1	8032.6	8032.6
Item	6	22174.5	3695.7
Subject	7	26251.6	3750.2
SOA*Subject	7	7586.7	1083.8
Item*Subject	42	4208.8	100.2
Residuals	0	0.0	

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 (F1 + F2, 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.

```
> 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

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

	lmer: p(t)	lmer: p(MCMC)	quasi-F	By-subject	By-item	F1+F2
Without treatment effect						
$\alpha = 0.05$	**0.088**	0.032	0.055	**0.310**	**0.081**	**0.079**
$\alpha = 0.01$	**0.031**	0.000	0.005	**0.158**	**0.014**	**0.009**
With treatment effect						
$\alpha = 0.05$		0.16	0.23			
$\alpha = 0.01$		0.04	0.09			

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

	lmer: p(t)	lmer: p(MCMC)	quasi-F	Subject	Item	F1 + F2
	Without treatment effect					
$\alpha = 0.05$	0.055	0.027	0.052	**0.238**	**0.102**	**0.099**
$\alpha = 0.01$	0.013	0.001	0.009	**0.120**	**0.036**	**0.036**
	With treatment effect					
$\alpha = 0.05$	0.823	0.681	0.809			
$\alpha = 0.01$	0.618	0.392	0.587			

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 1 error rate: 0.045 for $\alpha = 0.05$, 0.006 for $\alpha = 0.010$), whereas the mixed-effects model using the *t* test is on target (Type 1 error rate: 0.052 for $\alpha = 0.05$, 0.010 for $\alpha = 0.01$). Power is slightly greater for the mixed analysis evaluating probability using the *t*-test ($\alpha = 0.05$: 0.84 versus 0.81 for the quasi-F test; $\alpha = 0.01$: 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-

Table 6
Mean squares decomposition for the data with a Latin Square design in Raaijmakers et al. (1999)

	Df	Sum Sq	Mean Sq
Group	2	1696	848
SOA	2	46	23
List	2	3116	1558
Group∗Subject	9	47305	5256
SOA∗List	2	40	20
Residuals	18	527	29

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.

```
>summary(latinsquare.lmer4)
Random effects:
```

Groups	Name	Variance	Std.Dev.
Word	(Intercept)	754.542	27.4689
Subject	(Intercept)	1476.820	38.4294
Residual		96.566	9.8268

```
number of obs: 144, groups: Word, 12; Subject, 12
Fixed effects:
```

	Estimate	Std. Error	t value
(Intercept)	533.9583	13.7098	38.95
SOAmedium	2.1250	2.0059	1.06
SOAshort	−0.4583	2.0059	−0.23

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

```
> pvals.fnc(latinsquare.lmer4, nsim = 10000)
```

	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	Pr(>|t|)
(Intercept)	533.9583	534.0570	503.249	561.828	0.0001	0.0000
SOAmedium	2.1250	2.1258	−1.925	5.956	0.2934	0.2912
SOAshort	−0.4583	−0.4086	−4.331	3.589	0.8446	0.8196

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.

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

```
> latinsquare.lmer1 = lmer2(RT ~ SOA + (1|Word) + (1|Subject) + (1|Group) + (1 + SOA |List),
 data = latinsquare)
> latinsquare.lmer2 = lmer2(RT ~ SOA + (1|Word) + (1| Subject) + (1|Group) + (1|List), data =
 latinsquare)
> latinsquare.lmer3 = lmer2(RT ~ SOA + (1|Word) + (1 |Subject) + (1|Group), data = latinsquare)
> latinsquare.lmer4 = lmer2(RT ~ SOA + (1|Word) + (1 |Subject), data = latinsquare)
> latinsquare.lmer5 = lmer2(RT ~ SOA + (1|Subject), data = latinsquare)
>  anova(latinsquare.lmer1, latinsquare.lmer2, latinsquare.lmer3, latinsquare.lmer4,
latinsquare.lmer5)
```

	Df	AIC	BIC	logLik	Chisq Chi	Df	Pr(>Chisq)
latinsquare.lmer5.p	4	1423.41	1435.29	−707.70			
latinsquare.lmer4.p	5	1186.82	1201.67	−588.41	238.59	1	<2 e-16
latinsquare.lmer3.p	6	1188.82	1206.64	−588.41	0.00	1	1.000
latinsquare.lmer2.p	7	1190.82	1211.61	−588.41	1.379e-06	1	0.999
latinsquare.lmer1.p	12	1201.11	1236.75	−588.55	0.00	5	1.000

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.

($F(2,141) = 0.944$, $p = 0.386$) and the corresponding p-value calculated from the MCMC samples ($p = 0.391$). The mixed-effects analysis has slightly superior power

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 analyis (F1)

		Without SOA			With SOA		
		lmer: p(F)	lmer: p(MCMC)	F1	lmer: p(F)	lmer: p(MCMC)	F1
$\alpha = 0.05$	With	0.055	0.053	0.052	0.262	0.257	0.092
$\alpha = 0.01$	With	0.011	0.011	0.010	0.082	0.080	0.020
$\alpha = 0.05$	Without	0.038	0.036	0.043	0.249	0.239	0.215
$\alpha = 0.01$	Without	0.010	0.009	0.006	0.094	0.091	0.065

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

	Type I error rate				Power			
	Full		Missing		Full		Missing	
$\alpha = 0.05$	0.046	(0.046)	0.035	(0.031)	0.999	(0.999)	0.995	(0.993)
$\alpha = 0.01$	0.013	(0.011)	0.009	(0.007)	0.993	(0.993)	0.985	(0.982)

Evaluation based on Markov chain Monte Carlo sampling are listed in parentheses.

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 `latinsquare.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.

```
> 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)
```

	Df	AIC	BIC	logLik	Chisq Chi	Df	Pr(>Chisq)
dat.lmer4.p	5	9429.0	9452.4	−4709.5			
dat.lmer3.p	6	9415.0	9443.1	−4701.5	15.9912	1	6.364e-05
dat.lmer2.p	8	9418.8	9456.3	−4701.4	0.1190	2	0.9423
dat.lmerl.p	10	9419.5	9466.3	−4699.7	3.3912	2	0.1835

```
> print(dat.lmer3, corr = FALSE)
Random effects:
```

Groups	Name	Variance	Std.Dev.
items	(Intercept)	447.15	21.146
subjects	(Intercept)	2123.82	46.085
Residual		6729.24	82.032

```
Number of obs: 800, groups: items, 40; subjects, 20
Fixed effects:
```

	Estimate	Std. Error	t value
(Intercept)	362.658	16.382	22.137
listlistB	18.243	23.168	0.787
primingunprimed	31.975	10.583	3.021
listlistB: primingunprimed	−6.318	17.704	−0.357

The estimates are close to the parameters that generated the simulated data: $\sigma_i = 20$, $\sigma_s = 50$, $\sigma = 80$, $\beta_{\text{int}} = 400$, $\beta_{\text{priming}} = 30$, $\beta_{\text{list}} = 18.5$, $\beta_{\text{list:priming}} = 0$.

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

	$\beta_Z = 0$					
	$\alpha = 0.05$			$\alpha = 0.01$		
	X	Y	Z	X	Y	Z
lmerS: p(t)	0.609	0.990	**0.380**	0.503	0.982	**0.238**
lmerS: p(MCMC)	0.606	0.991	**0.376**	0.503	0.982	**0.239**
subj	0.677	0.995	**0.435**	0.519	0.979	**0.269**
item	0.210	0.873	0.063	0.066	0.670	0.012
lmer: p(t)	0.248	0.898	0.077	0.106	0.752	0.018
lmer: p(MCMC)	0.219	0.879	0.067	0.069	0.674	0.013
	$\beta_Z = 4$					
	$\alpha = 0.05$			$\alpha = 0.01$		
	X	Y	Z	X	Y	Z
lmerS: p(t)	0.597	0.989	0.925	0.488	0.978	0.867
lmerS: p(MCMC)	0.594	0.989	0.924	0.485	0.978	0.869
subj	0.650	0.992	0.931	0.487	0.979	0.868
item	0.183	0.875	0.574	0.055	0.642	0.295
lmer: p(t)	0.219	0.897	0.626	0.089	0.780	0.415
lmer: p(MCMC)	0.190	0.881	0.587	0.061	0.651	0.304

lmer: mixed-effect regression with crossed random effects for subject and item; lmerS: mixed-effect model with subject as random effect; Subj: by-subject regression; Item: by-item regression.

ticipant and item with random intercepts only. This model reflected exactly the structure implemented in the simulated data. A second mixed-effects model ignored the item structure in the data, and included only participant as a random effect. This model is the mixed-effects analogue to the by-participant regression.

Table 9 reports the proportions of simulation runs (on a total of 1000 runs) in which the coefficients of the regression model were reported as significantly different from zero at the 5% and 1% significance levels. The upper part of Table 9 reports the proportions for simulated data in which an effect of Z was absent, with $\beta_Z = 0$. The lower part of the table lists the corresponding proportions for simulations in which Z was present ($\beta_Z = 4$). The bolded numbers in the upper part of the table highlight the very high Type I error rates for models that ignore by-item variability that is actually present in the data. The only models that come close to the nominal Type I error rates are the mixed-effects model with crossed random effects for subject and item, and the by-item regression. The lower half of Table 9 shows that of these three models, the power of the mixed-effects model is consistently greater than that of the by-item regression. (The greater power of the by-subject models, shown in grey, is irrelevant given their unacceptably high Type I error rates.)

Of the two mixed-effects models, it is only the model with crossed random effects that provides correct estimates of the standard deviations characterizing the random effects, as shown in Table 10. When the item random effect is ignored (lmerS), the standard deviation of the residual error is overestimated substantially, and

Table 10
Actual and estimated standard deviations for simulated regression data

	Item	Subject	Residual
Data	40	80	50
lmer	39.35	77.22	49.84
lmerS		76.74	62.05

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, Bliese, 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 electrophysiology. In the past, journal policy in psychophysiological 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 psychophysiological 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 psychophysiologists 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 psycholinguists 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 Kliegl, 2007, and Kliegl, 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 $\boldsymbol{\Sigma}$ in Eq. (10). Interestingly, the profiled deviance function, which is negative twice the log-likelihood of model (10) evaluated at $\boldsymbol{\Sigma}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ for a given set of parameters, can be estimated without having to solve for $\boldsymbol{\beta}$ or $\mathbf{b}$. 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., σ_i/σ) on the horizontal axis, and the relative standard deviation of the subject random effect (σ_s/σ) 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 $\mathbf{Z}$. 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-

[2] 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.

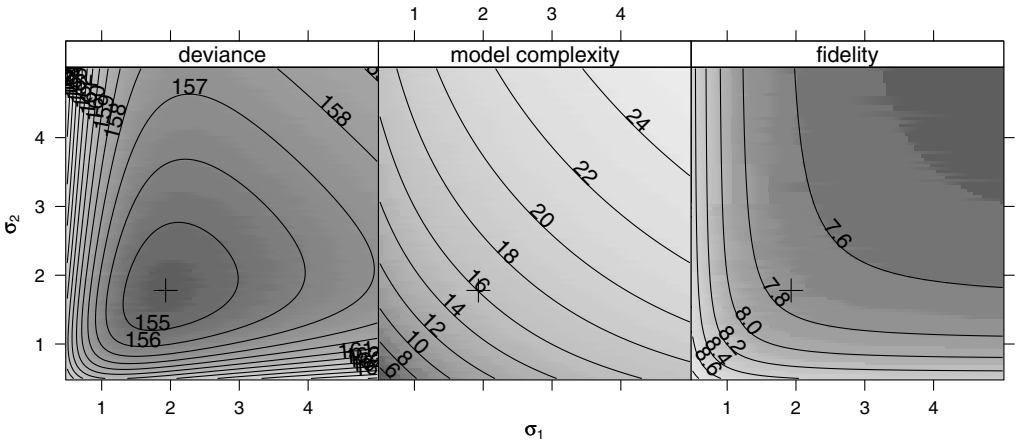

Fig. 2. Contours of the profiled deviance as a function of the relative standard deviations of the item random effects (x-axis) and the subject random effects (y-axis). The leftmost panel shows the deviance, the function that is minimized at the maximum likelihood estimates, the middle panel shows the 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.

mum likelihood estimates. The + symbols in each panel denote the values of the deviance components at the maximum likelihood estimates.

References

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modeling of data on teaching styles. *Journal of the Royal Statistical Society, A, 144*, 148–161.

Aitkin, M., & Longford, N. (1986). Statistical modeling in school effectiveness studies. *Journal of the Royal Statistical Society, A, 149*, 1–43.

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language, 40*, 41–61.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning, 50*, 5–43.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge: Cambridge University Press.

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology, 37*, 13–20.

Bates, D. M., & Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes*, R package version 0.99875-6.

Bates, D. M. (2005). Fitting linear mixed models in R. *R News, 5*, 27–30.

Bates, D. M. (2007). *Linear mixed model implementation in lme4*. Manuscript, university of Wisconsin - Madison, January 2007.

Bates, D. M. (2007a). *Fitting linear, generalized linear and mixed models to large data sets*. Paper presented at useR!2007, Ames, Iowa, August 2007.

Bedny, M., Aguirre, G. K., & Thompson-Schill, S. L. (2007). Item analysis in functional magnetic resonance imaging. *Neuroimage, 35*, 1093–1102.

Boyle, M. H., & Willms, J. D. (2001). Multilevel modeling of hierarchical data in developmental studies. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*, 141–162.

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. New York: Wiley.

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

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports, 14*, 219–226.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical models, design, and statistical model. *Annual Review of Psychology, 57*, 505–528.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Dien, J., & Santuzzi, A. M. (2004). Application of repeated measures ANOVA to high-density ERP datasets: A review and tutorial. In T. Handy (Ed.), *Event-related potentials*. Cambridge: MIT Press.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 package. *Journal of Statistical Software, 20*, 1–18.

Faraway, J. J. (2006). *Extending the linear model with R*. Boca Raton, FL: Chapman & Hall/CRC.

Felsen, G., & Yang, D. (2005). A natural approach to studying vision. *Nature Neuroscience, 8*, 1643–1646.

Fielding, A., & Goldstein, H. (2006). Cross-classified and multiple membership structures in multilevel models: An introduction and review. *Research Report No. 791. Department of Eduction and Skills*. University of Birmingham. ISBN 1 84478797 2.

Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed effect: Monte-Carlo estimates of error rates for F1, F2, F′, and minF′. *Journal of Verbal Learning and Verbal Behavior, 15*, 135–142.

Fox, J., Entink, R., & van der Linden, W. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software, 20*, 1–14.

Friston, K. J., Glaser, D. E., Henson, R. N. A., Kiebel, S., Phillips, C., & Ashburner, J. (2002a). Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage, 16*, 484–512.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-B., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping, 2*, 189–210.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., & Ashburner, J. (2002b). Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage, 16*, 465–483.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., & Kiebel, S. (2005). Mixed-effects and fMRI studies. *NeuroImage, 24*, 244âĂŞ–252âĂŞ.

Gardner, M. K., Rothkopf, E. Z., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory and Cognition, 15*, 24–28.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). AI, an efficient algorithm for REML estimation in linear mixed models. *Biometrics, 51*, 1440–1450.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J., & Thompson, R. (2002). *ASReml User Guide Release 1.0*. VSN International, 5 The Waterhouse, Waterhouse St, Hemel Hempstead, HP1 1ES, UK.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., et al. (1993). A multilevel analysis of school examination results. *Oxford Review of Education, 19*, 425–433.

Goldstein, H. (1995). *Multilevel statistical models*. London: Arnold.

Greenhouse, S. W., & Geisser, S. (1958). On methods in the analysis of profile data. *Psychometrika, 24*, 95–112.

Huttenlocher, J., Haight, W., Bryk, A., & Seltzer, M. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*, 236–249.

Jennings, J. R. (1987). Editorial policy on analysis of variance with repeated measures. *Psychophysiology, 24*, 474–478.

Kliegl, R. (2007). Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, & Reichle (2007). *Journal of Experimental Psychology: General, 136*, 530–537.

Kliegl, R., Risse, S., & 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.

Kreft, G. G. I. (1995). Hierarchical linear models: Problems and prospects. *Journal of Educational and Behavioral Statistics, 20*, 109–113.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 149–157.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Misangyi, V. F., LePine, J., Algina, J., & Goeddeke, F. (2006). The adequacy of repeated-measures regression for multilevel research. *Organizational Research Methods, 9*, 5–28.

MLwiN 2.1 (2007). Centre for Multilevel Modeling, University of Bristol, http://www.cmm.bristol.ac.uk/MLwiN/index.shtml.

Nutall, D., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research, 13*, 769–776.

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

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

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with the language-as-fixed-effect-fallacy: Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*, 416–426.

Raaijmakers, G. (2003). A further look at the language-as-a-fixed-effect fallacy. *Canadian Journal of Experimental Psychology, 57*, 141–151.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology, 52*, 501–525.

R development core team (2007). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, http://www.R-project.org.

Ruse, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience, 8*, 1647–1650.

Schroeder, M. L., & Hakstian, A. R. (1990). Inferential procedures for multifaceted coefficients of generalisability. *Psychometrika, 55*, 429–447.

Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and residual growth models. *Journal of Educational and Behavioral Statistics, 23*, 323–355.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. London: Sage Publications.

Vaan, L. de, Schreuder, R., & Baayen, R. H. (2007). Regular morphologically complex neologisms leave detectable traces in the mental lexicon. *The Mental Lexicon, 2*, 1–23.

Vasey, M. W., & Thayer, J. F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology, 24*, 479–486.

West, B. T., Welch, K. B., & Gałlechki, A. T. (2007). *Linear mixed models. A practical guide using statistical software*. Boca Raton: Chapman & Hall/CRC.

Willet, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10*, 395–426.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

Worsley, K. J., Liao, C., Aston, J., Petre, V., Duncan, G. H., & Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage, 15*, 1–15.