Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 59 (2008) 413–425
Journal of Memory and Language
www.elsevier.com/locate/jml

Examples of mixed-effects modeling with crossed random effects and with binomial data

Hugo Quené *, Huub van den Bergh

Utrecht Institute of Linguistics OTS, Utrecht University, Trans 10, NL-3512 JK Utrecht, The Netherlands

Received 28 February 2007; revision received 15 January 2008
Available online 24 April 2008

Abstract

Psycholinguistic data are often analyzed with repeated-measures analyses of variance (ANOVA), but this paper argues that mixed-effects (multilevel) models provide a better alternative method. First, models are discussed in which the two random factors of participants and items are crossed, and not nested. Traditional ANOVAs are compared against these crossed mixed-effects models, for simulated and real data. Results indicate that the mixed-effects method has a lower risk of capitalization on chance (Type I error). Second, mixed-effects models of logistic regression (generalized linear mixed models, GLMM) are discussed and demonstrated with simulated binomial data. Mixed-effects models effectively solve the "language-as-fixed-effect-fallacy", and have several other advantages. In conclusion, mixed-effects models provide a superior method for analyzing psycholinguistic data.

Keywords: Mixed-effects models; Crossed random effects; Analysis of variance; Logistic regression; GLMM

In most psycholinguistic studies, the experimenter presents multiple test items to multiple participants. Participants provide a response, which may be quasi-normally distributed (e.g. response times) or binomially distributed (e.g. yes/no responses). Typically, a particular sample of test items is presented to multiple participants, and a particular sample of participants responds to the same test items (often with rotation of test items over treatment conditions, to avoid repetition effects that would occur if each participant would be presented with all items in all conditions). For several decades, it has been common practice to analyze such data by means of two analyses of variance, following Clark (1973). The first analysis (yielding an F_1 for the treatment effect) has subjects or participants as a single random effect, whereas the second analysis (yielding F_2) has items as a single random effect. Clark (1973) also argued that these two F statistics for a fixed effect should be combined into a single minF' statistic. In current practice, this latter step is often ignored, in the belief that "if both F statistics are significant, [then] the effect is reliable over both subjects and items" (Raaijmakers, Schrijnemakers, & Gremmen, 1999, p. 419).

In this paper, we argue for the use of a new method of data analysis, called mixed-effects modeling, which is fully capable of performing a full analysis with multiple random factors simultaneously. This new method is also known as the multilevel model, hierarchical linear model, or variance component model. It has strong roots in biomedical and educational research, where researchers need

* Corresponding author.
E-mail address: hugo.quene@let.uu.nl (H. Quené).

0749-596X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2008.02.002

to acknowledge multiple random factors affecting their data: students' school performances are affected by the individual students, their school class, their school, etc. These random factors are typically in *nested* hierarchical order. A similar design occurs in language research if we attempt to model some phonetic property of a speech phrase, taking into account the phrases and their speakers as two nested random effects. The basic mixed-effects model needed for such designs, limited to *nested* random factors and applied to normally distributed data, was already introduced in a previous tutorial (Quené & van den Bergh, 2004), which serves as a background for the present paper (see also Faraway, 2006; Goldstein, 1999; Hoffmann & Rovine, 2007; Hox, 1995; Kreft & De Leeuw, 1998; van der Leeden, 1998; Luke, 2004; Maxwell & Delaney, 2004; Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002; Richter, 2006; Searle, Casella, & McCulloch, 1992; Snijders & Bosker, 1999).

In the present paper, our previous tutorial on mixed-effects models is extended into two directions. First, we introduce mixed-effects models in which the random effects are not nested but *crossed* (hence, we prefer the term "mixed-effects" over "multilevel" models for this paper). This is immediately relevant for psycholinguistic research, where participants and items typically constitute two crossed random effects, and researchers wish to generalize across both. We will explain why models with crossed random factors are superior to repeated-measures analysis of variance (RM-ANOVA), both from a theoretical perspective and in Monte Carlo simulations. In addition, we will demonstrate how to conduct these mixed-effects analyses in the statistical packages MLwiN and R, using response times from a phoneme monitoring study.

Second, we will extend the basic model with analysis methods for binary data, which follow a non-normal, binomial distribution. Applying the mixed-effects model to these data amounts to mixed-effects logistic regression, or applying a Generalized Linear Mixed Model (GLMM).

Mixed-effects modeling

The basic model for ordinary regression may be represented as:

$$Y_i = \beta_0 + \beta_1 X_i + (e_i) \quad (1)$$

The only random term here (in parentheses) is the residual e_i which represents the component of Y that is *not* captured by the regression equation. The predictor X is a fixed effect, which means that the values of β_0 and β_1 are assumed to be fixed throughout the population. The i observations are assumed to be sampled independently from the population, with uncorrelated and random deviations e_i.

In many situations, however, the i observations of Y are not independent, but they are grouped within higher-level units. For example, students are grouped into school classes, individuals are grouped into families, and psycholinguistic responses are grouped within participants. This group structure implies that the observations on the lower-level units (students, children, observations) are not independent but correlated—they depend to some extent on the particular higher-level units (schools, families, participants) sampled for the study. The basic mixed-effects model captures this dependency by adding a random effect u for each higher-level unit's unique contribution to the regression:

$$Y_{ij} = \gamma_0 + \gamma_1 X + (u_{0j} + e_{ij}) \quad (2)$$

where i indicates the lower-level units (e.g. observations), j indicates the higher-level units (e.g. participants), and subscript 0 indicates that, that term is constant for the corresponding units. In this model, the random effect u_{0j} represents the deviation of the j-th higher-level unit's average (participant average) from the overall intercept γ_{00}. Thus it captures the within-participant correlation among the lower-level units (observations).

The model for a repeated-measures analysis of variance also acknowledges this hierarchical structure (e.g. O'Brien & Kaiser, 1985; Pollatsek & Well, 1995; Raaijmakers et al., 1999; Rietveld & Van Hout, 2005). In RM-ANOVA, the random variance is separated into a variance component attributed to participants or subjects ($\sigma^2_{u_{0j}}$), and a variance component for within-participant residuals ($\sigma^2_{e_{ij}}$). Only the latter is used in tests of significance of the fixed effects. Note that the higher-level units have different intercepts, but identical slopes γ_1 in this model. This is equivalent to the assumption of homoscedasticity: the variance attributed to the fixed effect X is assumed to be identical for all higher-level units. In other words, the effect X is only found in the fixed part of the model, and not in the random part.

For several reasons, the mixed-effects model in (2) is to be preferred over a conventional RM-ANOVA (Baayen, Tweedie, & Schreuder, 2002; Quené & van den Bergh, 2004). First, the mixed-effects model is often more powerful than the univariate or multivariate approaches to RM-ANOVA, especially if the sphericity assumption is violated—as it often is in real data. One reason for this is that variances and covariances may be modeled explicitly, which allows for better capturing of heterogeneous variances and of correlation structure in the data. Second, the two-level hierarchical structure can easily be extended to more hierarchical levels of sampling, e.g. students within classes within schools, or observations within participants (hence, the term 'multilevel modeling'), which is not possible in RM-ANOVA models. Thirdly, both discrete and categorical predictors can be included in a single model, as will be illustrated

below. Finally, the mixed-effects analysis model is very robust against missing data, provided that data are missing at random. Hence, researchers do not need to impute missing data with debatable imputation methods.

In order to clarify the similarities and differences among these models, let us consider simulated data from a fictitious study into lip displacement (or some other articulatory parameter) which was measured under 3 different articulatory conditions (here named High, Zero and Low), with 12 repeated-measures in each condition. There were 24 participants in this fictitious study. The data were simulated to yield average scores of +0.2, 0, −0.2 in the High, Zero and Low conditions, respectively, with variance $s = 1$. The data were also simulated to violate both the sphericity assumption and the homoscedasticity assumption. In this fictitious experiment, the 12 repetitions are nested under participants, allowing a hierarchical or multilevel model.

In this paper, we will use two software modules for estimating and reporting mixed-effects models, each with its stronger and weaker points. The oldest of these is MLwiN (Rasbash et al., 2000), an incarnation of the older program `MLn` for DOS. The more recent one is the function `lmer` in package `lme4` (Bates, 2005). This is an extension package to R, an open-source package for statistical analysis available from http://www.r-project.org (R Development Core Team, 2008). The package `lme4` must be downloaded and installed separately. The advantages of MLwiN are that it offers more flexibility in specifying the random part of the model (and hence, in modeling heteroscedasticity and asphericity), because each term in the variance–covariance matrix can be specified explicitly, and second, that it offers a routine for evaluating user-defined contrasts in the fixed part of the model (command `ftest`). MLwiN is only available under Windows, however, and specifying a model with crossed random effects is somewhat awkward. The advantages of function `lmer` in R are that it is integrated in an excellent multi-platform software package, and second, that specifying crossed random factors is easy and straightforward. However, testing contrasts in the fixed part is more complicated in `lmer`. In the analyses reported below, the two programs' estimates are approximately equal. In order to familiarize readers with mixed-effects modeling, and with these two software tools, annotated logs of the analyses below are available from the first author, at http://www.let.uu.nl/~Hugo.Quene/personal/multilevel.

First, we examine the variance components of the higher-level and lower-level units, by fitting an empty model that does not contain any explanatory variables except for the grand mean or intercept (Snijders & Bosker, 1999):

$$Y_{ij} = \gamma_{00} + (u_{0j} + e_{ij}) \quad (3)$$

The results in Table 1 show a considerable intra-group variance of 0.277, corresponding to an intra-group correlation of $0.277/(0.277 + 0.800) = 0.27$. Hence, the hierarchical structure should not be ignored, because doing so would seriously inflate the risk of a Type I error (e.g. Quené & van den Bergh, 2004; Snijders & Bosker, 1999), even if the intra-group correlation were as low as .05.

A multivariate RM-ANOVA of the condition factor (within subjects), carried out in SPSS, yields $F(2, 22) = 4.325$, $p = .026$, with a reported effect size $\eta_p^2 = .282$ and estimated power of .69. However, this analysis falsely assumes that variances between subjects, and residual variances, are homogeneous among conditions.

The basic mixed-effects model in (2) contains a single continuous predictor X. If the predictor is categorical, however, it is more convenient to represent the three lev-

Table 1
Estimated parameters (with standard error of estimate in parentheses) of mixed-effects modeling of fictitious repeated-measures data

	Model (3)	Model (5)	Model (6)	Model (7)
Fixed				
γ_{00}	0.023 (0.116)		0.037	
γ_{H00}		0.216 (0.119)		0.216 (0.136)
γ_{Z00}		0.046 (0.119)		0.046 (0.108)
γ_{L00}		−0.191 (0.119)		−0.191 (0.146)
Random				
$\sigma^2_{u_{0j}}$	0.277 (0.086)	0.277 (0.086)		
$\sigma^2_{u_{H0j}}$			0.415 (0.137)	0.383 (0.127)
$\sigma^2_{u_{Z0j}}$			0.221 (0.080)	0.221 (0.080)
$\sigma^2_{u_{L0j}}$			0.508 (0.163)	0.455 (0.148)
$\sigma^2_{e_{ij}}$	0.800 (0.039)	0.771 (0.038)	0.696 (0.035)	0.696 (0.035)
Evaluation				
−2 log(lh)	2321.6	2290.9	2281.2	2277.2
χ^2_{deviance} (2 *df*)			40.4	13.7
p_{deviance}			<.001	.001
$\chi^2_{\text{condition}}$ (2 *df*)	n/a	31.25	n/a	9.44
$p_{\text{condition}}$		<.001		.009

els of this single factor as three dummy variables (High, Zero and Low), which have value 1 if the observation corresponds to that condition and value 0 otherwise. In this example, we could parameterize the three conditions as a baseline condition (Zero), and two treatment conditions (High and Low). Hence, the model has one term for the baseline or intercept, and two terms for the treatment effects (corresponding to the 2 degrees of freedom for this factor in a RM-ANOVA):

$$Y_{ij} = \gamma_{00} + \gamma_{H00}\mathrm{H} + \gamma_{L00}\mathrm{L} + (u_{0j} + e_{ij}) \quad (4)$$

Instead of this model, however, we prefer to suppress the overall intercept γ_0, and instead include the dummy variable for the Zero condition. The resulting coefficients, listed in Table 1, may now be interpreted directly as estimated means per condition:

$$Y_{ij} = \gamma_{H00}\mathrm{H} + \gamma_{Z00}\mathrm{Z} + \gamma_{L00}L + (u_{0j} + e_{ij}) \quad (5)$$

The two contrasts of interest are evaluated by comparing the estimated regression coefficients for the three dummy variables (taking the standard error of the estimate into account). These estimates for conditions 1 and 3 (see Table 1) differ by more than two standard errors. Hence, the null hypothesis is rejected on the basis of this significant difference. For more details on evaluation, see Faraway (2006), Goldstein (1999), Quené and van den Bergh (2004), Raudenbush and Bryk (2002), Snijders and Bosker (1999), Winer (1971). In MLwiN, the joint contrasts among the dummies' coefficients in the fixed part are evaluated by means of a χ^2 test statistic [here, $\chi^2(2) = 31.25, p < .001$], much like the single F test statistic obtained in a RM-ANOVA. In using `lmer` within R, fixed effects may be tested by means of the likelihood ratio tests outlined below, or by means of the function `aovlmer.fnc` in package `languageR`.

Both the RM-ANOVA and mixed-effects models above assume that the data are homoscedastic and spherical. In terms of model (5), this means that the participants differ in their individual intercept values u_{0j} but not in their regression coefficients γ_H, γ_Z, γ_L. The treatment effects are only found in the fixed part of the model. In this data set, however, the assumptions of homoscedasticity and sphericity are both violated. The effect of treatment conditions differs among participants, i.e., the variance between individual participants' averages is not the same for each treatment condition. If predictor X were continuous, then the slope of X would be different among participants.

This may be captured in the mixed-effects regression model, by including the dummy variables also in the random part of the model, at the higher (participant) level. We run two models, with and without the condition factor in the fixed part:

$$Y_{ij} = \gamma_{00} + (u_{Z0j}Z + u_{H0j}H + u_{L0j}L + e_{ij}) \quad (6)$$

$$Y_{ij} = \gamma_{Z00}\mathrm{Z} + \gamma_{H00}\mathrm{H} + \gamma_{L00}\mathrm{L} + (u_{Z0j}\mathrm{Z} + u_{H0j}\mathrm{H} + u_{L0j}\mathrm{L} + e_{ij}) \quad (7)$$

The estimated coefficients of model (6) (see Table 1) indicate that the between-subjects variances σ^2_{u0j} may not be homogeneous among treatment conditions. In the final model (7), differences between conditions in the fixed part are still significant [$\chi^2(2) = 9.44$, $p = .009$], although the χ^2 test statistic is considerably smaller than in the previous model (5). The incorrect assumption of the sphericity in that model (5), i.e. ignoring the true asphericity in the data, has greatly inflated the significance of the treatment effect, and may well have led here to capitalization on chance.

Although the tables in the present paper report standard errors for the random estimates (following e.g. Kreft & De Leeuw, 1998; Snijders & Bosker, 1999), these are now widely regarded as unsafe for evaluating random estimates. Faraway (2006, chap. 8) and Baayen, Davidson, and Bates (2008) discuss several methods for evaluating fixed and random effects. We have chosen to evaluate random terms in the various models by means of likelihood ratio tests, which compare the likelihood values of these models. If two models are identical in their fixed parts, and if one model is completely contained within the other, and if both models are based on maximum likelihood estimation, then their difference in likelihood values may be evaluated by means of a χ^2 distribution of the deviance (difference in likelihood), with the number of additional parameters in the more detailed model as the degrees of freedom (Faraway, 2006; Hox, 1995; Pinheiro & Bates, 2000; Snijders & Bosker, 1999). Hence, model (6) is compared against model (3), and the final model (7) is compared against model (5), with 2 *df* The latter evaluation shows that the final model (7) fits the data significantly better than did the predecessor model (5) [$\chi^2(2) = 13.7$, $p = .001$]. Because the latter mixed-effects model also captures the random variances due to violations of homoscedasticity and sphericity, it performs significantly better than an ANOVA-like model that (often incorrectly) assumes these violations to be absent.

Crossed random effects

The fictitious data analyzed in the preceding section came from some hierarchical sampling design, in which the repeated-measures were *nested* within participants; our previous tutorial (Quené & van den Bergh, 2004) was also limited to such hierarchical designs. Crucially, the repetitions were not correlated across participants. A similar multilevel, hierarchical structure is encountered in many quasi-experimental studies, where the lower-level units are usually uncorrelated across higher-level units. In a corpus study of spoken Dutch, for example, we investigated the speaking rate of phrases (lower-level units) that were spontaneously produced by speakers (higher-level

units) (Quené, 2008). This is a typical hierarchical sampling structure: phrases are nested under speakers, and phrases are uncorrelated across speakers, since each speaker spontaneously produced an almost unique sample of phrases from the population of possible phrases.

In well-controlled psycholinguistic experiments, however, the repeated observations within the higher-level units (participants) are typically performed using the same sample of test items across participants. In terms of the experiment, this means that the observed results may be limited to the selected test items only, and that we cannot safely generalize results to other possible test items (Clark, 1973). In more statistical terms, this means that the residuals e_{ij} in (2) are still correlated, so that these errors are in fact not independent. In other words, the experimental design contains two random effects that are crossed and not nested. There is usually only a single observation for each combination of participant j and test item k. Analyzing such data with RM-ANOVA is problematic, because RM-ANOVA allows for only one such random effect in its model. As proposed by Clark (1973), two independent analyses are performed, each with one random effect (either participants, yielding F_1, or test items, yielding F_2).

The basic multilevel models above can be extended to allow for multiple random effects that are crossed, and not nested (Goldstein, 1999; Snijders & Bosker, 1999). Such models with crossed random effects are ideally suited to analyse results from our psycholinguistic experiments, because they allow for simultaneous and joint generalization to other participants and to other test items. This follows from the simultaneous inclusion of both random factors into a single analysis. Moreover, the aforementioned general advantages of mixed-effects modeling (no assumptions of homoscedasticity nor sphericity, robustness against missing data, mixing discrete and continuous predictors) also apply to models with crossed random effects. In short, this type of models provides a superior alternative to the current practice of splitting the problem into two separate RM-ANOVAs over participants and over items.

The empty mixed-effects model for crossed random effects is given as:

$$Y_{i(jk)} = \gamma_{0(00)} + \left(u_{0(j0)} + v_{0(0k)} + e_{i(jk)}\right) \tag{8}$$

This model contains three terms in its random part: the unique component of each participant $u_{0(j0)}$, of each test item $v_{0(0k)}$, and the residual component $e_{i(jk)}$. (The latter corresponds to the deviation of each observation from its predicted value.)

In order to illustrate this model, let us return to the same fictitious data analyzed before. In this section, we regard these responses as coming from a fictitious experiment with three presentation conditions (e.g. in sentence contexts with high, neutral, and low semantic predictability of the target word). The three conditions constitute a *fixed* factor, which varies systematically and not randomly, and all values of which are present in the experiment in roughly equal proportions. A sample of 36 suitable target words was distributed over the three conditions, with 12 target words or test items in each condition. Items were rotated over conditions and participants, so that each participant responded to each of the 36 test items only once. Conversely, each test item was presented to 24 participants, with eight participants in each condition, according to a latin square (yielding an incomplete design; Cochran & Cox, 1957).

This fictitious data set (identical to the one analyzed in the preceding section) was generated by assuming a crossed design, with heteroscedasticity and asphericity. This was achieved by using different variances and covariances in the three treatment conditions at the levels of participants and of items. At the participant level, the variance–covariance matrix used in generating values of u was $\begin{bmatrix} 0.200 & & \\ 0.245 & 0.300 & \\ 0.283 & 0.346 & 0.400 \end{bmatrix}$. At the items level, the single variance used in generating values of v was $\sigma^2_v = 0.2$ for all treatment conditions. At the residual level (nested under both participants and items), the variance–covariance matrix used in generating values of e was $\begin{bmatrix} 0.6 & & \\ & 0.5 & \\ & & 0.4 \end{bmatrix}$. Responses were assumed to follow the normal distribution.

The results of the empty model (8) in Table 2 show that the total variance in the random part is now decomposed into three components, viz. variance between participants $\sigma^2_{u_{0(j0)}} = 0.288$, variance between test items $\sigma^2_{v_{0(0k)}} = 0.257$, and residual variance nested under the combination of participants and test items $\sigma^2_{e_{i(jk)}} = 0.540$. Since a substantial amount of the random variance is due to participants and to test items, both of these random variances should be included in the model. (An intra-group correlation coefficient cannot be reported here because it is not defined for models with crossed random effects.)

The final, optimal model contains the dummy variables for the three conditions in the fixed part, and also in the random part at the participant level:

$$\begin{aligned} Y_{i(jk)} = {} & \gamma_H \mathrm{H} + \gamma_N \mathrm{N} + \gamma_L \mathrm{L} \\ & + \left(u_{H0(j0)}\mathrm{H} + u_{N0(j0)}\mathrm{N} + u_{L0(j0)}\mathrm{L} + v_{0(0k)} + e_{i(jk)}\right) \end{aligned} \tag{9}$$

Just like model (7) above, this current optimal model captures the heteroscedasticity and asphericity in the data, and also captures random effects of participants and of items simultaneously. This final model (9) there-

Table 2
Estimated parameters (with standard error of estimate in parentheses) of mixed-effects modeling of fictitious data from a study with two crossed random factors

	Model (8)		Model (9)	
Fixed				
$\gamma_{0(00)}$	0.023 (0.141)		0.042 (0.105)	
$\gamma_{H0(00)}$		0.216 (0.145)		0.216 (0.142)
$\gamma_{N0(00)}$		0.046 (0.145)		0.046 (0.137)
$\gamma_{L0(00)}$		−0.191 (0.145)		−0.191 (0.156)
Random				
$\sigma^2_{u_{0(j0)}}$	0.288 (0.089)	0.289 (0.089)		
$\sigma^2_{u_{H0(j0)}}$			0.344 (0.114)	0.306 (0.103)
$\sigma^2_{u_{N0(j0)}}$			0.264 (0.090)	0.267 (0.091)
$\sigma^2_{u_{L0(j0)}}$			0.461 (0.148)	0.400 (0.130)
$\sigma^2_{v_{0(0k)}}$	0.257 (0.067)	0.258 (0.066)	0.209 (0.056)	0.212 (0.056)
$\sigma^2_{e_{i(jk)}}$	0.540 (0.027)	0.510 (0.025)	0.495 (0.025)	0.495 (0.025)
Evaluation				
−2 log(lh)	2080.7	2034.8	2087.0	2082.2
χ^2_{deviance} (2 *df*)			6.3	47.4
p_{deviance}			.043	<.001
$\chi^2_{\text{condition}}$ (2 *df*)	n/a	47.23	n/a	5.05
$p_{\text{condition}}$		<.001		.080

Intermediate untitled models are also reported for comparison.

fore provides a far better fit [−2(lh) = 2082.2] than the corresponding non-crossed model (7) [−2(lh) = 2277.2].

The current model is more conservative for testing fixed effects than the corresponding non-crossed models discussed above. This follows from the partitioning of the total variance in a fixed and a random part. The random part of model (7) is ameliorated in (9), by including terms for $u_{0(j0)}$ and $v_{0(0k)}$ simultaneously (and by including different terms for the between-subject variances in different treatment conditions). Consequently, the random part now explains a larger amount of the total variance, so that the variance components for the fixed terms in the model have to decrease. This is indicated by the larger standard errors in the fixed part for model (9) in Table 2, as compared to those for model (7) in Table 1. Stated otherwise, underestimating the random variances by ignoring random effects inflates the risk of a Type I error regarding the fixed effects. The main effect of condition is not significant in the final crossed model above.

The conventional two-step RM-ANOVA on these data yields $F_1(2,22) = 4.325$, $p = .026$, $F_2(2,34) = 14.826$, $p < .001$, $\min F'(2,34) = 3.348$, $p = .047$. The joint contrast for the fixed effects in model (7) was reported above as $\chi^2(2) = 9.44$, $p = .009$, whereas the joint contrast in the final model (9) is $\chi^2(2) = 5.05$, $p = .080$, exceeding $\alpha = .05$. According to this final model, then, the null hypothesis should in fact *not* be rejected. As expected, the mixed-effects model is indeed more conservative than the other analysis methods (RM-ANOVAs with evaluation of F_1 and F_2, RM-ANOVAs with evaluation of joint $\min F'$, mixed-effects modeling without the item effect, and mixed-effects modeling ignoring heteroscedasticity).

Monte Carlo simulations

In order to compare the outcomes of fully specified, crossed mixed-effects models with the conventional RM-ANOVA outcomes more systematically, we have generated 500 simulations of the data set. Simulation parameters for treatment effects, heteroscedasticity and asphericity were similar to those used in the fictitious data analyzed above. These 500 data sets were analyzed in three ways: first, by computing F_1 and F_2 using multivariate analysis (thus correcting for asphericity), second, by computing $\min F'$, and third, by fitting a mixed-effects model (similar to (9) above), all at $\alpha = .05$. For comparison, we have also generated and analyzed 500 simulated data sets *without* a fixed effect of condition, i.e. according to the null hypothesis. These Monte Carlo outcomes are summarized in Fig. 1.

If conditions are indeed different, as in the lower panel of Fig. 1, then the power in detecting this fixed effect (1) using joint F_1 and F_2 (all outcomes in the upper right quadrant) was .950, (2) using $\min F'$ was .902, and (3) using crossed mixed-effects model was .886. Thus the latter analysis is indeed slightly more conservative than the ANOVA-based methods.

The justification of this more conservative character of the crossed mixed-effects model is found in the upper panel of Fig. 1, here showing the probability of incorrectly rejecting the null hypothesis. This probability of Type I error in detecting the condition effect (1) using joint F_1 and F_2 (all outcomes in the upper right quadrant) was .140, (2) using $\min F'$ was .064, and (3) using crossed mixed-effects model was .048,

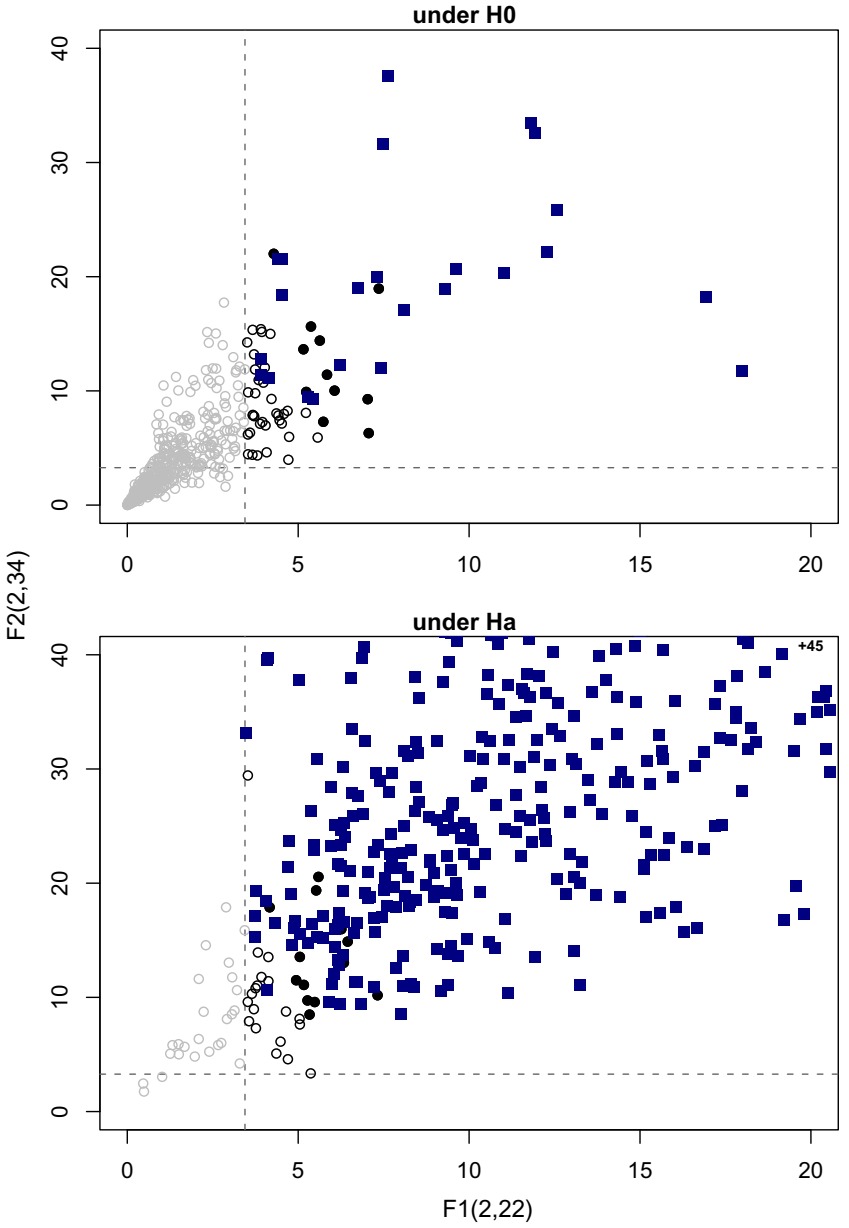

Fig. 1. Resulting F_1 and F_2 values of simulated data sets, with indication of significance of the fixed effect of condition, according to joint F_1 and F_2 (open circles), minF' (filled circles), and crossed mixed-effects model (squares). Upper panel: without fixed effect of condition (according to H_0; 500 outcomes); lower panel: with fixed effect of condition (according to H_a; showing 455 of 500 outcomes). Dashed lines indicate critical F ratios.

close to the nominal $\alpha = .05$. These outcomes support the findings of Raaijmakers et al. (1999), who argue that the current practice of reporting only F_1 and F_2 is dangerous because it inflates the risk of a Type I error. The present simulations confirm this inflation of Type I error. Although many researchers regard minF' as too conservative for comfort, these simulations suggest that it may even be too liberal, because it ignores the joint correlation within subjects and within items. This point is further addressed below.

Application to response time data

In order to further demonstrate the mixed-effects modeling with crossed random effects, we will analyze response time data obtained from a phoneme monitor-

ing experiment. These response times were obtained in a Dutch replication of the parallel English study by Quené and Port (2005), which itself was a modification of a similar study by Pitt and Samuel (1990). The 36 participants were instructed to press a response button as soon as they heard a pre-specified target phoneme in a list of spoken disyllabic words (Connine & Titone, 1996). The critical factor was whether the word list had regular timing (isochronous intervals between stressed syllables) or irregular timing (non-isochronous intervals); faster responses are predicted for the regularly timed stimuli as compared to the irregular-timed stimuli. The stress pattern of the target word, being either trochaeic (strong–weak) or iambic (weak–strong), constituted a second factor in this study. As a third factor, the target phoneme occurred either in the first or in the second syllable.

Hence, there are three fixed effects, viz. regularity (within items, within participants), stress pattern, and target position (both varying between-items, within-participants). For each of the four cells defined by the latter two factors, 24 target words were selected, with an appropriate target phoneme in the appropriate position. The 4×24 Dutch items were rotated over the presentation conditions, so that a participant heard a particular target words only once, yielding four different versions of the experiment (with nine participants per version). Response times were measured from the onset of the target phoneme. The individual responses were stored in univariate format, i.e. each row of the data file corresponds to a single response, with participant, item, factor values (and additional information) coded in each row. RTs were transformed to their logarithmic values, in order to remove the intrinsic positive skew and non-normality of their distribution (Keene, 1995; Limpert, Stahel, & Abbt, 2001). There were 198 missing responses (6%); these were simply discarded and not imputed.

As above, the preliminary mixed-effects model contains only the intercept or "grand mean" in its fixed part. In the random part (in parentheses), the total variance is decomposed into three components, viz. variance between participants $\sigma^2_{u_{0(j0)}}$, variance between test items $\sigma^2_{v_{0(0k)}}$, and residual variance nested under the combination of participants and test items $\sigma^2_{e_{i(jk)}}$.

$$\log(Y_{\mathrm{i(jk)}}) = \gamma_{0(00)} + \left(u_{0(\mathrm{j}0)} + v_{0(0\mathrm{k})} + e_{\mathrm{i(jk)}}\right) \tag{10}$$

The estimated coefficients for this model are shown in Table 3. The variance between items $\sigma^2_{v_{0(0k)}}$ is considerably larger than the variance between participants $\sigma^2_{u_{0(j0)}}$, as expected, because two of the factors of interest vary between test items. Hence, the between-items variance in the empty model reflects systematic effects under investigation.

The next step is to include the fixed factors of interest to the model. All three factors are binary (having two levels), and they are entered as a single dummy factor, with value 1 or "on" indicating timing being regular (hence, TR), stress falling on the second syllable (i.e., a trochee target word, hence, S2), and the target phoneme being positioned in the second syllable (hence, P2), respectively. Because of this parameterization of terms in the model, the main effect of timing regularity models this effect in the condition with trochee target words (dummy for stress equals zero) with the target phoneme in the first syllable (dummy for position equals zero). Relevant two-way interactions are also included. (The two-way interaction between regularity and target position was never significant, and it has been eliminated here for clarity.) The estimated coefficients are also listed in Table 3. Note that the random variance between items is greatly reduced, now that a considerable part of this variation has been captured in the fixed part of the regression model. Each main effect is significant, because its t value (estimate divided by standard error) even exceeds the critical value $t^*_{.95} = 2.03$ for 36 df.

Next, it should be verified whether the data are indeed homoscedastic, as assumed by this regression model. This was done by including each main effect into the random part at both the levels of participants and of items, and evaluating all these candidate models by means of their likelihood values, as outlined above. Based on this exploration, the optimal model contains the effect of regularity both in the fixed part and in the random part at the items level:

$$\begin{aligned}\log\left(Y_{\mathrm{i(jk)}}\right) = {} & \gamma_{0(00)} + \gamma_{\mathrm{TR}0(00)}\mathrm{TR} + \gamma_{\mathrm{S2}\ 0(00)}\mathrm{S2} \\ & + \gamma_{\mathrm{P2}\ 0(00)}\mathrm{P2} + \gamma_{\mathrm{TR:S2}\ 0(00)}\mathrm{TRS2} \\ & + \gamma_{\mathrm{S2:P2}\ 0(00)}\mathrm{S2P2} \\ & + \left(u_{0(\mathrm{j}0)} + v_{0(0\mathrm{k})} + v_{\mathrm{TR}\ 0(0\mathrm{k})}\mathrm{TR} + e_{\mathrm{i(jk)}}\right)\end{aligned} \tag{11}$$

The estimated coefficients in Table 3 for this final model (11) show that the main effect of regularity is no longer significant, if this effect is also included in the random part of the model. In other words, the between-items variability is considerably larger in the regular-timing condition than in the irregular-timing condition. If this heteroschadasticity is taken into account, then items with initial stress do not show a significant effect of regularity any more (due to the larger standard error of the estimate; $t = -1.79$). For items with final stress, however, the regularity effect is still highly significant ($t = -3.81$). As before, we see that ameliorating the random part, to account better for random variation, lowers the significance levels for the fixed effects, thus reducing the risk of capitalization on chance.

The results of this study involving lists of spoken words confirm that regular speech rhythm facilitates spoken-word perception, yielding faster reaction

Table 3
Estimated parameters (with standard error of estimate in parentheses) of mixed-effects modeling of response times obtained in a phoneme monitoring experiment

	Model (10)		Model (11)
Fixed			
$\gamma_{0(00)}$	6.269 (0.037)	6.338 (0.036)	6.338 (0.038)
$\gamma_{TR\ 0(00)}$		−0.040 (0.016)*	−0.041 (0.023)
$\gamma_{S2\ 0(00)}$		−0.275 (0.032)*	−0.281 (0.036)*
$\gamma_{P2\ 0(00)}$		0.245 (0.031)*	0.245 (0.030)*
$\gamma_{TR:S2\ 0(00)}$		0.121 (0.023)*	−0.121 (0.032)*
$\gamma_{S2:P2\ 0(00)}$		−0.015 (0.043)	−0.002 (0.042)
Random			
$\sigma^2_{u_{0(j0)}}$	0.028	0.028	0.0279
$\sigma^2_{v_{0(0k)}}$	0.052	0.008	0.0149
$\sigma^2_{v_{TR\ 0(0k)}}$			0.0121
$\sigma^2_{e_{i(jk)}}$	0.110	0.107	0.1035
Evaluation			
−2 log(lh)	2446	2188	2156
χ^2_{deviance} (1 *df*)			32
p_{deviance}			<.001

* $p < .05$.

times—but only for iambic words, and not for trochee words. This matches with similar findings for American English, reported and discussed by Quené and Port (2005). Hence, the present results support the notion that in American English and Dutch, which are rhythmically quite similar, the perceptual effects of timing regularity on spoken-word recognition are also quite similar.

Binomial data

In our previous tutorial covering basic mixed-effects modeling (Quené & van den Bergh, 2004), and in our exposition above, we have assumed that the response variable follows a normal distribution—perhaps after some monotonous data transformation to normalize responses. Mixed-effects modeling is also available, however, for data from a binomial distribution, such as binary responses in a psycholinguistic study. Binomial data are typically analyzed by means of logistic regression (Hosmer & Lemeshow, 2000; Kleinbaum, 1992; Pampel, 2000). Hence, the present extension may be regarded as a mixed-effects version of logistic regression models, or Generalized Linear Mixed Models (GLMM) (Cnaan, Laird, & Slasor, 1997; Guo & Zhao, 2000).

The current practice in many studies seems to be that the raw responses (hits and misses) are aggregated at a higher-level (of subjects or of items), and the resulting counts or proportions (e.g. hit rates or miss rates) are then fed into a conventional RM-ANOVA. For example, hit rates are computed per condition per participant, aggregating over items, and these aggregated proportion data are then analyzed in RM-ANOVA with condition as a within-subject factor. Because counts and proportions are not normally distributed, this practice violates one of the assumptions underlying ANOVA. Moreover, the practice of aggregating over a random factor also reduces the error variance used in further tests of significance (e.g. Quené & van den Bergh, 2004; Snijders & Bosker, 1999), thus inflating the risk of a Type I error (i.e., of capitalization on chance). Mixed-effects logistic models provide a more conservative and balanced method of analyzing binomial data.

In a binomial data set, the variance is defined as a function of the mean hit rate π: $\sigma^2 = \pi(1-\pi)$, so the mean and variance are related. Because of this non-independence, mixed-effects modeling of binomial data is technically complex (Goldstein, 1999; Snijders and Bosker, 1999). In particular, the u terms in the random part need to "mirror" the γ terms in the fixed part, in order for the model to yield sensible results: $\sigma^2_u = \gamma(1-\gamma)$ (Snijders & Bosker, 1999). Now that powerful implementations are widely available in `lmer` for R, and in the MLwiN program, researchers need hardly worry about such complications in mixed-effects logistic regression.

In order to explain mixed-effects logistic modeling, we have degraded the continuous and normally distributed response variable analyzed in (9) into a discrete binomial variable. Values of $Y < 1$ are regarded as 'miss' (0), and larger values are regarded as 'hits' (1). (One could imagine, for example, that the continuous variable represents some level of mental activation, and that the discrete variable represents the presence or absence of a behavioral response.) For the High, Neutral and Low conditions, the respective hit rates happen to be 0.24,

0.16, and 0.11. In logistic modeling, the hit rates are first converted to logit units [i.e. to the logarithm of the odds of hits: $\text{logit}(P) = \log(P/(1-P))$], which are then regressed onto the predictors in the model. Here the respective hit rates correspond to logit values of −1.174, −1.660, and −2.044, respectively. Note that negative logit values correspond to hit rates below .5.

As before, we start by fitting the empty model. Instead of aggregating over participants, or over items, however, we model the raw hits themselves. This means that we can also specify a model with crossed random factors, i.e., the logistic equivalent of model (8):

$$\text{logit}\left(Y_{\text{i(jk)}}\right) = \gamma_{0(00)} + \left(u_{0(\text{j0})} + v_{0(0\text{k})}\right) \tag{12}$$

Because of the computational complexities outlined above, it is not possible to estimate variance components for residuals, as these are implied by the mean(s). Because we have specified crossed random factors, the intra-group correlation cannot be reported. The resulting estimates are listed in Table 4.

The subsequent model has the condition effect in its fixed part only. The difference with conventional RM-ANOVA is that both participants and items are again included as two crossed random factors:

$$\text{logit}\left(Y_{\text{i(jk)}}\right) = \gamma_{\text{H0(00)}}\text{H} + \gamma_{\text{N0(00)}}\text{N} + \gamma_{\text{L0(00)}}\text{L} + \left(u_{0(\text{j0})} + v_{0(0\text{k})}\right) \tag{13}$$

The joint contrasts among the dummies' coefficients in the fixed part yield a significant main effect: $\chi^2 = 14.96$, $p < .001$, which suggests significant differences in hit rates among the three conditions. The conventional RM-ANOVA reports a (marginally) significant main effect in both analyses [$F_1(2,22) = 3.2$, $p = .058$; $F_2(2,34) = 6.3$, $p = .005$] although the minF' value is not significant [min$F'(2,43) = 2.13$, $p = .131$].

As before, we verify whether the assumption of homoscedasticity is warranted, by evaluating candidate models having the predictors in the random part. This shows that the between-subject variance $\sigma^2_{u_{0(j0)}}$ varies among the treatment conditions, much like the original continuous variable did (cf. Model (9)):

$$\text{logit}\left(Y_{\text{i(jk)}}\right) = \gamma_{\text{H0(00)}}\text{H} + \gamma_{\text{N0(00)}}\text{N} + \gamma_{\text{L0(00)}}\text{L} + \left(u_{\text{H0(j0)}}\text{H} + u_{\text{N0(j0)}}\text{N} + u_{\text{L0(j0)}}\text{L} + v_{0(0\text{k})}\right) \tag{14}$$

The estimated coefficients for this final model are also listed in Table 4. The far larger variance between participants in the Low condition is now represented correctly in the random part, and not in the fixed part. This results in a larger standard error (less confidence) in the estimate for this condition in the fixed part, which in turn results in the main effect being not significant [$\chi^2(2) = 4.39$, $p = .111$]. As before, an artefactually significant effect in the fixed part (i.e., a Type I error) has disappeared when the random part was ameliorated. Although the two models (13) and (14) perform equally well [as indicated by their non-significant deviance, $\chi^2(2) = 3.3$, $p = .192$], we prefer the latter model, because it better captures the lower confidence for very low (or very high) hit rates.

Discussion and conclusion

The exposition above has demonstrated that mixed-effects modeling provides an excellent tool for analyzing

Table 4
Estimated parameters (with standard error of estimate in parentheses) of mixed-effects modeling of fictitious binomial data (see text for details)

	Model (12)	Model (13)	Model (14)
Fixed			
γ_{000}	−1.585 (0.256)		
γ_{H000}		−1.174 (0.279)	−1.174 (0.260)
γ_{N000}		−1.660 (0.290)	−1.660 (0.273)
γ_{L000}		−2.045 (0.304)	−2.045 (0.405)
Random			
$\sigma^2_{u_{0(j0)}}$	0.943 (0.332)	0.968 (0.340)	
$\sigma^2_{u_{H0(j0)}}$			0.830 (0.380)
$\sigma^2_{u_{N0(j0)}}$			0.715 (0.393)
$\sigma^2_{u_{L0(j0)}}$			2.807 (1.055)
$\sigma^2_{v_{0(0k)}}$	0.642 (0.223)	0.653 (0.227)	0.612 (0.219)
Evaluation			
−2 log(lh)	652.7	611.7	615.0
χ^2_{deviance} (2 *df*)			3.3
p_{deviance}			.192
$\chi^2_{\text{condition}}$ (2 *df*)	n/a	14.96	4.39
$p_{\text{condition}}$		<.001	.111

data from psycholinguistic experiments. Multiple random factors can be included into a single analysis, which essentially solves the "language-as-fixed-effect-fallacy" that has plagued researchers for decades (Clark, 1973; Meuffels & van den Bergh, 2006). The multiple random factors can be nested (e.g. observations within participants, for repeated-measures designs with independent observations across participants), or they can be crossed (e.g. items and participants, for repeated-measures designs where observations are also correlated across participants). Well-established techniques are available for modeling data that are normally distributed (perhaps after some data transformation, as illustrated above) and for binomial data. Models for other distribution are also under current development.

Only a few years ago, mixed-effects modeling was a complicated operation requiring specialized software, which often imposed a steep learning curve. So far this has prevented mixed-effects modeling from becoming standard routine in behavioral linguistic research, contrary to other disciplines. Since powerful and convenient implementations have now become available, however, running one mixed-effects model is often easier than running two RM-ANOVAs. Hence, there is little reason to maintain the conventional RM-ANOVA routine of computing and combining F_1 and F_2.

The results above also indicate that mixed-effects models are typically more conservative than conventional RM-ANOVA, at least for designs with crossed random effects. In our Monte Carlo simulations above, the power of the fixed effect of condition was .90 for analyses using minF', and .89 for crossed mixed-effects analyses. In other studies the reduction of power may not be so mild, depending on the actual experimental design, and the degrees of heteroscedasticity, of asphericity, and of intra-group correlation (cf. Kreft & De Leeuw, 1998, chap. 5.4). [For nested designs, however, multilevel analysis seems to be *more* powerful than a single RM-ANOVA, Quené and van den Bergh (2004)]. The conservative behavior of crossed mixed-effects models is nevertheless appropriate, as the following metaphor illustrates. Let us regard the data set under analysis as a sponge, soaked with information. The water in the sponge represents the variance in the data set. Our research hypothesis is concerned with the true variance (water) related to the fixed effects in our study. Hence, we have to squeeze out of the sponge as much of the random variance as we can, leaving only the fixed variance within the sponge, and we catch that squeezed-out random variance in a bucket. In RM-ANOVA, we are limited to use only one hand for squeezing. So, we squeeze with the right hand (yielding F_1), we take an exact copy of the soaked sponge, and we squeeze that with the left hand (yielding F_2). In both squeezes, however, some amount of random variance (water) is inadvertedly left in the sponge, because squeezing can only be done imperfectly with a single hand. In both analyses, this remaining random variance weighs in with the fixed-effect variance in the sponge, instead of with the random variance in the bucket. Hence, to mix metaphors, it inflates the significance of the fixed effect, and may result in a false positive outcome, or Type I error. In mixed-effects modeling, by contrast, we use both hands for squeezing. This allows us to squeeze more random variance (water) out of the sponge. After squeezing, the amount of variance (water) within the sponge is smaller, and the amount in the bucket is larger, because of our more effective two-handed squeezing. Because the amount of random variance (water) that is left in the sponge instead of the bucket is now smaller, the evaluation of the fixed effects will be better. The less liberal behavior of mixed-effects analysis, relative to minF' analyses, provides a better protection against capitalization on chance.

If the language items under study (e.g. target words) are regarded as a random sample from a larger population of these items, i.e. as a random effect, then crossed mixed-effects models effectively solve the "language-as-fixed-effects-fallacy", as argued before. However, including items as a random effect in a crossed mixed-effects model does not in itself provide information about the generality of the fixed effect(s) over these items (as noted by Leonard Katz in the internet-based debate about F_2; see Forster, this issue, for an introduction). Such a measure of generality would even go against this model, precisely because items are assumed to be randomly selected, and hence, as essentially homogeneous in their susceptibility to the fixed effect. Moreover, an interaction of items by treatments cannot be evaluated in an incomplete design such as the latin-square designs discussed here. Unless all combinations of subjects and items are tested under all treatment conditions, the two-way interactions of items by treatments and of subjects by treatments are confounded with the main effect of treatment (Bailey, 1982; Cox, 1958; Janse & Quené, 2004). What analysts *can* do, however, is inspect and compare the residuals of the items (cf. Afshartous & Wolf, 2007), and attempt to relate these to some relevant property of the items. This property (e.g. part of speech, lexical frequency, etc.) may then be included as a tentative *fixed* between-item predictor in their candidate models, and it may be tested for its main effect and for interactions with the fixed treatment effects under study. In an educational analogy, we can evaluate the fixed effect of treatment on students' performance, but we cannot evaluate the interaction of treatment and students, if students were indeed sampled at random. Relevant properties of the students, e.g. IQ score, may however be used as additional fixed predictors of the treatment effect, and of the variation in students' individual intercepts and in students' slopes of the treatment effect.

Mixed-effects modeling thus offers considerable advantages over RM-ANOVA techniques, as discussed above. This technique also allows researchers to com-

bine discrete and categorical predictors into a single model, it does not require homoscedasticity nor sphericity, it does not require aggregation of binomial data, and it is robust against missing data. In conclusion, mixed-effects modeling provides a superior means of statistical analyses in our discipline, and adopting this technique is highly recommended to improve our research into language and speech behavior.

Acknowledgments

Hugo Quené and Huub van den Bergh, Utrecht institute of Linguistics OTS, Utrecht University.

Our thanks are due to Douglas Bates for technical assistance, to Esther Janse and Sieb Nooteboom for helpful discussions, and to Tom Lentz, Harald Baayen and two anonymous reviewers for valuable comments and suggestions.

References

Afshartous, D., & Wolf, M. (2007). Avoiding 'data snooping' in multilevel and mixed effects models. *Journal of the Royal Statistical Society Series A (Statistics in Society), 170*(4), 1035–1059.

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

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language, 81*(1–3), 55–65.

Bailey, R. (1982). Confounding. In N. Johnson & S. Kotz (Eds.). *Encyclopedia of statistical sciences* (Vol. 2, pp. 128–134). New York: Wiley.

Bates, D. (2005). Fitting linear models in R: Using the `lme4` package. *R News, 5*(1), 27–30.

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

Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine, 16*(20), 2349–2380.

Cochran, W., & Cox, G. (1957). *Experimental designs*. New York: Wiley.

Connine, C. M., & Titone, D. (1996). Phoneme monitoring. *Language and Cognitive Processes, 11*(6), 635–645.

Cox, D. (1958). *Planning of experiments*. New York: Wiley.

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: Chapman and Hall.

Goldstein, H. (1999). *Multilevel statistical models* (2nd (electronic update) ed.). London: Edward Arnold.

Guo, G., & Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology, 26*, 441–462.

Hoffmann, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*(1), 101–117.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Hox, J. (1995). *Applied multilevel analysis*. Amsterdam: TT Publicaties.

Janse, E., & Quené, H. (2004). On measuring multiple lexical activation using the cross-modal semantic priming technique. In H. Quené & V. J. van Heuven (Eds.), *On speech and language: Studies for Sieb G. Nooteboom* (pp. 105–114). Utrecht: LOT.

Keene, O. N. (1995). The log transformation is special. *Statistics in Medicine, 14*, 811–819.

Kleinbaum, D. G. (1992). *Logistic regression: A self-learning text*. New York: Springer.

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

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Lognormal distributions across the sciences: Keys and clues. *Bioscience, 51*(5), 341–352.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Meuffels, B., & van den Bergh, H. (2006). De ene tekst is de andere niet: The language-as-fixed-effect fallacy revisited: Statistische implicaties. *Tijdschrift voor Taalbeheersing, 28*(4), 323–345.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin, 97*(2), 316–333.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.

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

Pitt, M. A., & Samuel, A. G. (1990). The use of rhythm in attending to speech. *Journal of Experimental Psychology: Human Perception and Performance, 16*(3), 564–573.

Pollatsek, A., & Well, A. D. (1995). On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(3), 785–794.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of the Acoustical Society of America, 123*(2), 1104–1113.

Quené, H., & Port, R. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica, 62*(1), 1–13.

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*(1–2), 103–121.

R Development Core Team. (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing. [Computer program (version 2.6.2)]. Available from: http://www.r-project.org.

Raaijmakers, J. G., Schrijnemakers, J. M., & 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.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., et al. (2000). A user's guide to *MLwiN* (Computer program (version 2.02) and manual. Available from: http://www.cmm.bristol.ac.uk/team/userman.pdf). Multilevel Models Project, Institute of Education, University of London.

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

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41*(3), 221–250.

Rietveld, T., & Van Hout, R. (2005). *Statistics in language research: Analysis of variance*. Berlin: Mouton de Gruyter.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

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

van der Leeden, R. (1998). Multilevel analysis of repeated measures data. *Quality and Quantity, 32*, 15–19.

Winer, B. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.