Behav Res (2012) 44:232–247
DOI 10.3758/s13428-011-0145-1

Twice random, once mixed: Applying mixed models to simultaneously analyze random effects of language and participants

Dirk P. Janssen

Published online: 22 August 2011

Abstract Psychologists, psycholinguists, and other researchers using language stimuli have been struggling for more than 30 years with the problem of how to analyze experimental data that contain two crossed random effects (items and participants). The classical analysis of variance does not apply; alternatives have been proposed but have failed to catch on, and a statistically unsatisfactory procedure of using two approximations (known as F_1 and F_2) has become the standard. A simple and elegant solution using mixed model analysis has been available for 15 years, and recent improvements in statistical software have made mixed models analysis widely available. The aim of this article is to increase the use of mixed models by giving a concise practical introduction and by giving clear directions for undertaking the analysis in the most popular statistical packages. The article also introduces the DJMIXED add-on package for SPSS, which makes entering the models and reporting their results as straightforward as possible.

Keywords ANOVA · Mixed models · Hierarchical linear modeling · Item effects · Language as a fixed effect fallacy

Electronic supplementary material The online version of this article (doi:10.3758/s13428-011-0145-1) contains supplementary material, which is available to authorized users.

D. P. Janssen
Department of Psychology, University of Kent,
Canterbury, Kent, UK

Present Address:
D. P. Janssen (✉)
International Media and Entertainment Management,
NHTV University of Breda,
Mgr. Hopmansstraat 1,
4817 JT, Breda, The Netherlands
e-mail: dirk.janssen@gmail.com

When the design of an experiment that is to be analyzed with an analysis of variance (ANOVA) is considered, there is a fundamental statistical difference between fixed factors and random factors. An informal definition of random factors is that they test only a subset of all possible levels of that factor and that there are no theoretical implications of the outcomes at each level of the factor. Participants are the most common random factor in psychology experiments. Only a subset of all available participants is tested, and there is little or no theoretical interest in the performance of individual participants (of course, the individual participants should be inspected and screened, and if unexpected patterns of participant behavior are found, this should be indicated).

Coleman (1964) and Clark (1973) realized that in all language experiments, there are two random factors: participants and words. One could argue that this realization came too early for its own good: At that point in time, there was no fully satisfactory way to deal with two random factors in one ANOVA. Clark's suggestion of using *F'* or *minF'* failed to catch on, despite evidence of its virtues (Forster & Dickinson, 1976; Santa, Miller, & Shaw, 1979; Wickens & Keppel, 1983; and others). In his article, Clark also provided an alternative technique of using two approximate values, which was intended for reanalyzing existing experiments for which the raw data were no longer available. This method of using two approximate values, F_1 and F_2, has become the de facto standard in the psycholinguistic literature.

Briefly refreshing well-known concepts will hopefully aid understanding (but the impatient reader can skip ahead to Example 1). The F test is a ratio of two variance estimates. It is the between-conditions variance divided by the within-conditions variance. The between-conditions variance is an estimate of the influence of the factor under scrutiny. The within-conditions variance (the *error term*) is an estimate of

the noise in the data, at least in the simple case. We assume a significant effect of the condition if the F test indicates that the between-conditions variance is sufficiently larger than the error term, where sufficiency is determined by the F distribution and the degrees of freedom.

The presence of two random factors causes the estimate of between-conditions variance to be contaminated by unwanted interactions. For a design with exactly one random factor, the unwanted interactions can be canceled out by choosing an error term that also contains those interactions. However, for a design with two random factors, no appropriate error term exists (Clark, 1973; see also Baayen, Tweedie, & Schreuder, 2002; Jackson & Brashers, 1994; Keppel & Wickens, 2004, p. 539). Hence, no algebraically exact F test can be computed for these designs, if one stays within the framework of the ANOVA.

The solutions for this problem that have been proposed before in the psycholinguistic literature can be divided into four strands. First, some have proposed using the F' (or *minF'*) test, a test that constructs an approximate error term (Forster & Dickinson, 1976; Maxwell & Bray, 1986; Santa et al., 1979; Wickens & Keppel, 1983). Second, some have proposed doing two analyses, F_1 and F_2. In each of these analyses, one random factor is analyzed, and the other is treated as fixed, leading to two approximate tests that should then be evaluated together (Wike & Church, 1976). Despite its obvious shortcoming of being based on incorrect assumptions, this method has become standard practice. Third, some have argued that, at least in certain experimental settings, items could be classified as a fixed factor, by-passing the problem all together (Raaijmakers, Schrijnemakers, & Gremmen, 1999).

This article favors a fourth solution that is based on *mixed modeling*, a technique for combining random and fixed factors into one analysis, which has been developed since the 1980s. The name *mixed modeling* refers to mixing random and fixed effects, but the same technique is also known under other names, such as *cross-classificational hierarchical linear models* (Raudenbush, 1993). This technique bypasses all the problems that the classical ANOVA has, making it possible to present a very simple and straightforward analysis of a design containing two random factors.

This fourth solution has a number of separate origins, most of which are outside of psycholinguistics. The mixed, crossed-effects model that is the key to this solution was used in Raudenbush (1993) and was further extended by Rasbash and Goldstein (1994). An independently created application to item response theory (IRT), including predictors at the participant level, has been proposed by Mislevy (1987) and others (see also Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003, on the relationship between IRT and mixed models). Baayen et al. (2002) were probably the first to publish on it in the context of psycholinguistic experimentation.

Mixed models and hierarchical models have been popularized in a number of recent publications: For example, Cheng, Sheu, and Yen (2009) have demonstrated the use of subject-specific random effects in the expectancy valence model of the Iowa gambling task. For the analysis of eye-tracking data, Barr (2008) has made a case for using hierarchical logistic regression. Lee and Vanpaemel (2008), Rouder and Lu (2005), and Shiffrin, Lee, Kim, and Wagenmakers (2008) have written on the general applicability of hierarchical Bayesian methods for building and comparing models of cognitive tasks. The aim of the present article is to popularize the mixed model for psychologists dealing with language materials.

Whether a factor is treated as random or fixed is, to a certain extent, up to the researcher (Jackson & Brashers, 1994). The choice has implications for the generalizability of the findings, for the type of statistical questions that can be asked, for the fit between data and the model, and for the conceptual match between the model and the theory. There are various ways to define random factors (or more generally speaking, random effects or random coefficients). In the hierarchical linear modeling (HLM) literature, random effects are introduced to account for the relatively high correlation between data points that fall within one hierarchical level (i.e., students in one class are more alike than students across classes). From a Bayesian perspective, random factors are introduced to increase generalizability and accuracy. From a classical (frequentist) point of view, random factors allow for more parsimonious models with fewer parameters.

Without a good knowledge of the statistical underpinnings, it may be hard to determine which factors are best treated as random in a psycholinguistic experiment. A helpful guideline is whether the levels that are tested are of direct theoretical relevance. Consider an experiment comparing the efficacy of complex and simplex primes in a lexical decision task. The levels of the fixed factor *prime type* (simplex and complex) are determined by and directly related to the theoretical question under consideration. The random factor *words* is different: Even if words are carefully selected to match on many dimensions, the actual words that are selected for the two levels of prime type do not affect the conclusions drawn from the experiment. In addition, there is no theoretical interest in the (average) priming effect exhibited by individual words, whereas the average priming effects of the prime type levels will be the main finding of the experiment. As a rule of thumb, one can look at the reported means: Studies invariably include a table listing the average reaction time (RT) for each level of each fixed factor in the experiment, whereas levels of the random factors are not reported or are delegated to an

appendix (see also Jackson & Brashers, 1994, for an in-depth discussion of random factors).

As defendants of the third solution on how to combine random items and participants, Raaijmakers et al. (1999) suggested analyzing items as fixed in those experiments that basically deplete the pool of possible words. In that case, they argued, items are not selected *at random*. However, random selection of words is a sufficient but not a necessary condition for treating items as a random factor (Jackson & Brashers, 1994; Raudenbush & Bryk, 2002). Even when (nearly) all possible items are used because of multiple selection restrictions, the choice of stimuli has little or no theoretical repercussions, and items should be treated as random to allow generalization of the findings beyond the set of items used in the experiment. Raaijmakers et al.'s argument would hold for a hypothetical experiment in phonology, in which the pronunciation time for the words *goat* and *goal* are compared with that for *coat* and *coal*. Here, the theoretical interest lies in the pronunciation times for these actual four words, and items can be treated as fixed.

The textbook ANOVA is limited to one random factor because of restrictions in its underlying linear model and the way this underlying model is computed. Another limitation of the textbook ANOVA is that fixed factors and random factors are treated very similarly, despite their apparent differences. The mixed model analysis overcomes both limitations: It allows for more than one random factor in the design, and it treats random factors inherently differently from fixed factors. The underlying linear model of the mixed model is different and can often be solved only by computer-intensive iteration and approximation techniques. To a researcher using a modern computer, this technical difference is irrelevant. Mixed models have additional benefits, such as that they can naturally handle unequal numbers of observations in the cells of the design.

Mixed models are closely related to HLMs. The emergence of hierarchical linear modeling has transformed statistical practice in many areas of the social sciences over the past 15 years (Raudenbush & Bryk, 2002) and has been available, in rudimentary form at first, in SPSS since version 11. Mixed models differ from HLMs in that they do not require a *hierarchical* relationship between the factors. In fact, HLM can be viewed as a special case of mixed modeling. This makes mixed models well suited for language research: It is hard to convincingly argue that *items* are nested under *participants* or that *participants* are nested under *items* (but see Richter, 2006, for a defense of the former). In almost all experiments, items and participants are crossed, since every participant will see each item or every participant will see one variant of each item, as has been previously argued by Baayen, Davidson, and Bates (2008), Quené and van den Bergh (2008), and others.

The fundamental difference between a textbook linear model (also called *classical linear model* or *ordinary least-squares linear model*) and a mixed model is the presence of random effects in the model. To see how random effects are represented, I will first revisit the representation of fixed effects.

Consider the following simplistic standard linear model of the simple experiment outlined above:

$$Y_j = \beta_0 + \beta_1 \cdot \text{PrimeType}_j + \varepsilon_j.$$

Here, PrimeType is dummy coded (simplex is zero and complex is one), and j indexes the different observations. This formula says that the observed RT can be modeled as the sum of the intercept (β_0), an influence of the variable PrimeType quantified by β_1, and error. Because of the dummy coding chosen, the value of PrimeType is zero for simplex words, so the intercept (β_0) will be the *expected* RT for all simplex words. For the complex words, PrimeType equals one, and an additional value (β_1) allows complex words to have a different expected RT from simplex words: The expected RT for complex words is $\beta_0 + \beta_1$. The error term ε_j is taken from a normal distribution with a mean of zero, which allows actual observations to differ from predictions.

If the experiment had four conditions, instead, and those conditions were treated as a fixed factor, the formula would look like this:

$$Y_j = \beta_0 + \beta_1 \cdot \text{PTB}_j + \beta_2 \cdot \text{PTC}_j + \beta_3 \cdot \text{PTD}_j + \varepsilon_j.$$

Here, the four levels of PrimeType are dummy coded in three variables, PTB to PTD. The variables β_1 to β_3 will represent the RT differences between levels B to D and the comparison level A, as shown in Table 1.

This is a variant of the items-as-a-fixed-effect model. There are two statistical complications with this model. First, when the number of levels of a fixed factor is increased, the model becomes more complex. Second, because the β terms are model parameters (they determine the content of the statistical model), any conclusions drawn from this experiment are, strictly speaking, conditional upon the values of β that were observed. This is less surprising than it may sound.

Table 1 Dummy coding for a predictor with four levels and corresponding expected reaction times, $E(Y)$

	Dummy			
PrimeType	PTB	PTC	PTD	$E(Y)$
A	0	0	0	β_0
B	1	0	0	$\beta_0 + \beta_1$
C	0	1	0	$\beta_0 + \beta_2$
D	0	0	1	$\beta_0 + \beta_2$

If a simplex versus complex difference of 100 ms was observed, the conclusion drawn from the first model would be that future research will also find a 100 ms difference. If the observed difference (β_1) is different, the prediction changes. Therefore, the predictions of the model are conditional on the values of the model parameters.

It follows that if the factor item with k levels (items) is modeled as a fixed factor, $k-1$ dummy variables and $k-1$ corresponding βs are included in the model formula. When interactions between item and another dummy-coded (categorical) factor R are included in the model, another $(k-1)\times(r-1)$ terms are required (where r is the number of levels of R). Clearly, applying the fixed factor approach to items can lead to very complex models: If there are 32 items in four conditions, $(32-1)\times(4-1)=93$ different β terms are required. Despite its complexity, this model would not constrain the values of the various βs at all. An experimenter would reasonably expect that items within one condition have similar average RTs (βs), but no statistical property of this items-as-a-fixed-effect model enforces this.

Whereas β_0 and β_1 are parameters of the model that represent one single value (e.g., a 500 ms intercept, a 100 ms condition effect), the error term ε is a vector parameter that represents as many values as there are observations: There is one value ε_j for each observation j. In statistical output, the individual values ε_j are normally not listed, but the variance of all values is reported as the variance of ε, denoted s_ε^2 or error variance. Mathematically, the values of the vector ε are modeled by a known statistical distribution with a certain mean and variance. The usual assumptions for ε are that its shape is that of the normal (Gaussian) distribution with a mean of zero and a variance that is a model parameter, the error variance s_ε^2. In terms of conditional inference, this means that the conclusions drawn from this model are conditional only on the variance of the error term and on the fact that the errors should be normally distributed, and not on the actual error values that were observed. In other words, the conclusions from one experiment hold for all future experiments in which a similar error variance is obtained.

Vector parameters are used in mixed models to model random effects in a way that is quite similar to the treatment of residuals in the classical model. Instead of assuming a different β for each level of the factor item, the levels are modeled by a vector parameter u, which has a different value for each item i. A value u_i reflects the relative speed of item i, as compared with the prototypical item in that condition. For an average item, u_i is close to zero, since the average item is close to the prototypical item. For a slow item, u_i is relatively large and positive, increasing the expected RT. For a fast item, u_i will be relatively large and negative, reducing RT.

In other words, the item-specific value u_i adjusts the expected RT to reflect the relative speed of item i. The conceptualization of u as a vector of *item-specific adjustments* to the modeled RTs has been shown to greatly aid understanding of mixed models. Quite intuitively, all adjustments u_i are assumed to center around a mean according to a normal distribution with a certain variance, as shown in Fig. 1. In the left panel, each line represents the effect of priming on one (hypothetical) item. The priming effect is constant, but some items are inherently faster or slower than others, which leads to a distribution of lines. In the right panel, the distribution of item adjustments around the overall item mean is shown, which is close to normal. This means that the items within one condition are modeled as necessarily similar to each other, with the majority of the items having an adjustment that is close to zero. Outlier items are possible, but they should be less likely the further they are removed from zero.

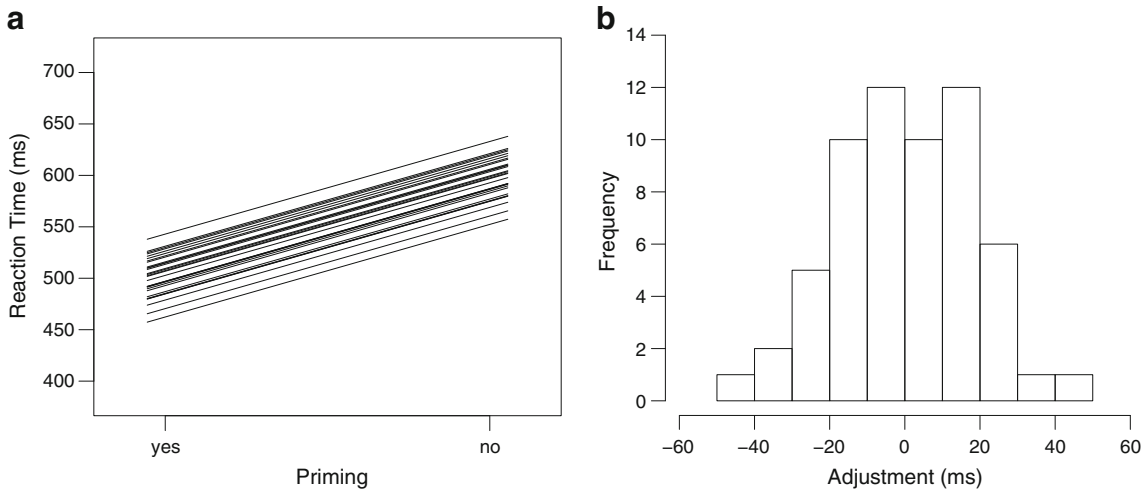

Fig. 1 Left panel: reaction times for 60 hypothetical items in unprimed and primed conditions; each item has an individual adjustment (u_i) to reflect its inherent speed. Right panel: distribution of adjustment values u_i is close to normal

Returning to the experiment with simplex and complex words, the expected RT for a morphologically complex item i will have three parts in a mixed model: the intercept β_0, the effect of condition β_1, and the *adjustment* specific to this item u_i. This leads to the following mixed model formula (which does not yet include any effects related to participants):

$$Y_{ij} = \beta_0 + u_i + \beta_1 \cdot \text{PrimeType}_i + \varepsilon_{ij}$$
$$\text{with} \quad u \in \mathcal{N}(0, \sigma_u^2) \quad \text{and} \quad \varepsilon \in \mathcal{N}(0, \sigma_\varepsilon^2).$$

Here, the jth observed RT for item i is modeled as the intercept β_0, a random effect (relative adjustment) for this ith item that has strength u_i, an effect of the PrimeType (simplex vs. complex) that has strength β_1, and a residual value specific to this observation. The values of vector u are taken from a normal distribution ($\mathcal{N}$) with a mean of zero and variance σ_u^2, making u similar in many ways to residual vector ε, which is taken from a normal distribution with a mean of zero and variance σ_ε^2. Because the average values of u_i and of ε_{ij} are both zero (see Fig. 1), the *expected* RT for any item is based on the intercept and the effect of PrimeType only:

$$E(Y) = \beta_0 + \beta_1 \cdot \text{PrimeType}.$$

Although a separate value u_i for each item is computed, only one model parameter is used. This model parameter represents the variance between items, s_u^2. The parametrization on the variance implies that the conclusions drawn from this model are conditional only on the observed variance between the words and the fact that they are approximately normally distributed with a mean of zero. In practical terms, the conclusions drawn from this experiment should hold for all future experiments in which a similar item variance is obtained. The items-as-a-fixed-effect model above would be conditional on the actual effects found for individual words.

Modeling the factor item with a random effect u in a mixed model has a number of interesting conceptual implications, when compared with the item-as-a-fixed-effect model discussed above: Similar to the earlier guideline criterion for random factors, modeling the values as a distribution in the mixed model agrees with a limited theoretical interest in specific values for each item. Second, the generalizability of the model with a random effect is greater, because the conclusions are conditional only on the variance of all words (s_u^2), and not on the number and values of the individual β weights, as in the items-as-a-fixed-effect case. In the mixed model, the length of the vector u is not a model parameter, whereas the number of β weights in the fixed effects model changes if more items are added. Because the effects of individual words are modeled as taken from a normal distribution, the mixed model assumes that most words are similar, centered around a prototypical or ideal word, with outlier words becoming less frequent as they get further removed from the average. If item is modeled as a fixed effect, there are no constraints on the values of the individual item effects (β weights) at all. Because a mixed model estimates the model means and parameters using a precision weighted average (see Raudenbush & Bryk, 2002, p. 40, for details), the number of observations per participant and per item can vary substantially without any repercussions for the analysis. Finally, as will be outlined in more detail below, testing the mixed model does not depend on approximating F-values, which allows for a larger and more diverse set of statistical questions that can be answered.

One note of caution is in place here: Statistical inference from one model to a second set of data is, strictly speaking, conditional on the actual value of each parameter of the model: An item-as-a-fixed-factor model based on 20 items cannot be extended to a data set with 21 items. In practice, researchers would use inductive reasoning to argue that extending the model to the second data set is reasonable: The similarity between the two data sets and the lack of evidence indicating that one extra item may drastically change the outcomes will convince most readers that this is appropriate. The critical consideration here is that a mixed model approach does not require inductive logic in this case, since the number of items is not a model parameter and the model can be applied without reservations.

We have so far dealt only with the effect of the random factor of items. Below, the example will be extended to include a random factor of participants. A commonly raised question is how this mixed model analysis (and especially one with a random factor for participants) compares with a repeated measures ANOVA. The difference between the two approaches is that in mixed models, items, participants, and the random variation between individual items and participants are included predictors in the model. A repeated measures analysis, on the other hand, includes only a predictor that captures the differences between participants' average RTs. With this, it can partition the sum of squares into three parts: within-condition variance, between-condition variance, and between-participant variance. The additional third source of variance sets it apart from the normal ANOVA and reduces the error term (within-condition variance) so that a more sensitive F value can be computed.

Example 1

After this theoretical and conceptual overview of how mixed modeling works and how it accounts for random

factors in the data, a number of examples will be presented that will be analyzed in SPSS. Both the standard SPSS MIXED syntax and the use of the SPSS extension package DJMIXED will be discussed. Matching syntax for SAS and the free statistical package *R* is supplied in the online Appendix.

The procedure and implications of using a mixed model analysis will be demonstrated from an example data set containing priming data obtained from 34 participants. Each participant made a lexical decision on 62 experimental items, for a total of 2,004 valid data points (5% missing data). The two factors of interest were priming (the critical word was the first word of a pair, priming absent, or the second word of a pair, priming present) and morph (the critical word was part of an inflectional or a derivational pair). Other properties of the items that might influence the RT were matched. All stimulus words were part of 31 triplets formed by a base word, one of its inflections, and one of its derivations. Each participant first saw either a derivation or an inflection, followed by the matching base word. As is shown in Table 2, there is an indication of an interaction between priming and morph, but the standard deviations of the cell means are sizable. (This is a real data set, in which I artificially strengthened the effect of morph for didactic purposes.)

For comparison, an F_1/F_2-based analysis, using repeated measures in the F_1, resulted in the following mixture of significances. The factor priming is significant by F_1 and by F_2, $F_1(1, 33) = 74.6$, $MSE = 2{,}609$, $p = .000$; $F_2(1, 30) = 77.4$, $MSE = 2{,}317$, $p = .000$. The factor morph is significant by F_1 and by F_2, $F_1(1, 33) = 18.3$, $MSE = 1{,}013$, $p = .000$; $F_2(1, 30) = 14.9$, $MSE = 1{,}204$, $p = .001$. The interaction between morph and priming is significant in F_1, but not in F_2, $F_1(1, 33) = 9.4$, $MSE = 860$, $p = .004$; $F_2(1, 30) = 3.9$, $MSE = 2{,}200$, $p = .058$.

Below, arguments for using a slightly more complex approach will be outlined, but it is instrumental to see what a very straightforward mixed model for these data looks like. In the terminology introduced above, the mixed model will contain item-specific *adjustments* to the predicted RTs, to model that some items are easy (negative adjustments) and some items are hard (positive adjustments). The set of all item-specific adjustments is modeled by a normal distribution with a mean of zero, which has two consequences: The average adjustment is zero, and larger adjustments should be less frequent than small adjustments.

In addition to the item-specific adjustments, we will also introduce participant-specific adjustments in this model. These adjustments are drawn from a second, independent normal distribution, and they model that some participants are fast (large negative adjustments) and some participants are slow (large positive adjustments), but most participants are close to average (have an adjustment that is close to zero).

We have to extend the notation introduced above to incorporate item-specific adjustments (u_{0i}) and participant-specific adjustments (u_{0p}). The letters i and p indicate the type of adjustment; the zero will be used later. The resulting model is identical to the corresponding classical regression or ANOVA model, but for the inclusion of the two random effects.

Table 2 Data for Example 1: average reaction times (in milliseconds) and standard deviations

		Morph	
		Derivation	Inflection
Priming	Unprimed	683 (172)	646 (162)
	Primed	594 (141)	588 (140)
Difference		89	58

Mixed model 1

A mixed model was fitted to the data that contained the fixed effects of priming, morph, and their interaction and two random effects accounting for participant-specific and item-specific adjustments to the intercept. The mixed model formula is

$$Y_{pi} = \beta_0 + u_{0p} + u_{0i} + \beta_1 \text{Priming}_i + \beta_2 \text{Morph}_i + \beta_3(\text{Priming}_i \times \text{Morph}_i) + \varepsilon_{pi}.$$

This model claims that observed RTs can be modeled with a general intercept term β_0, which is modified by a participant-specific adjustment u_{0p} (which distinguishes fast from slow participants) and an item-specific adjustment u_{0i} (which distinguishes fast from slow items). The expected RT is further modified by the effect of priming (of strength β_1), the effect of morph (of strength β_2), and their interaction (of strength β_3). Finally, the observed RTs differ from the predicted RTs by an observation-specific amount of error, ε_{pi} (error is observation specific because each participant sees each item only once).

Because both priming and morph are dummy coded, the interaction effect β_3 applies only to the observations for which morph equals 1 and priming equals 1. The presence of a significant interaction term tells us that the *combined* effect of morph and priming is different from the sum of their effects ($\beta_1 + \beta_2$).

The results of this model are summarized in Table 3. The fixed effect of priming was highly significant, $F(1, 72) = 47.9$, $p = .000$, the effect of morph reached significance, $F(1, 184) = 9.2$, $p = .003$, and so did the interaction, $F(1, 174) = 4.2$, $p = .042$. Both random effects were significant:

Table 3 Fixed and random effects for model 1, example 1

		Fixed Effects			
Model term		Category	β	F	p
Priming		Unprimed	59.55	47.9	.000
Morph		Inflected	8.29	9.2	.003
Priming × Morph		Unpr + Infl	32.44	4.2	.042
		Random effects			
Model term		Adjustment for	Variance	Z	p
u_{0p}	Intercept	Participants	5,144	3.90	.000
u_{0i}	Intercept	Words	1,774	4.27	.000
ε	Error	–	16,880	30.55	.000

for u_{0p}, $Z = 3.90$, $p = .000$; for u_{0i}, $Z = 4.27$, $p = .000$, showing that their inclusion was warranted.

One could clearly argue that a mixed model analysis is easier to report and easier to understand than the matching F_1/F_2 analysis, since there are simply fewer F tests. The significance of the random effects u_{0i} and u_{0p} should be reported, but they are not of direct theoretical relevance. The significance of u_{0i} and u_{0p} effectively means that items differed from each other and participants differed from each other, which is to be expected in any experiment. In fact, absence of significance should be discussed in more detail: If the effect of u_{0i} were nonsignificant, this could mean that items are almost identical, which is unexpected and may be theoretically interesting.

To stay within the classical ANOVA report, two degrees of freedom were reported for each F-test. However, for mixed models, the denominator degrees of freedom does not correspond to the number of cases or items, but it is computed in a different way (via the Satherthwaite method). The numerator degrees of freedom is identical to the number of levels of the factor minus one, as usual.

Although the model formula may look complex, the matching SPSS syntax is very simple when the SPSS extension module that was specifically written for this article is used (the extension module can be downloaded from the journal's archives and from djmixed.googlecode.com). Figure 2 shows the graphical interface to the DJMIXED package, while Fig. 3 shows the corresponding DJMIXED syntax: The part that spans lines 2–7 defines the mixed

Fig. 2 DJMIXED point-and-click interface

```
1 * SPSS syntax when using the extension package.
2 DJMIXED /MIXEDMODEL
3             DV = rt
4             PREDICTORS = priming morph priming*morph
5             PPS = Participant
6             ITEMS = Word
7             NAME = 'interaction'.
8 * Produces a short summary.
9 DJMIXED /MODELSUMMARY
10            NAME = 'interaction'.
```

Fig. 3 DJMIXED syntax for Example 1

```
1 * SPSS syntax for Model 1: fixed effects and interaction.
2 MIXED rt BY priming morph
3 /FIXED= priming morph priming*morph
4 /RANDOM=intercept | SUB(Participant) COVTYPE(VC)
5 /RANDOM=intercept | SUB(Word) COVTYPE(VC)
6 /METHOD=ML
7 /PRINT=COVB SOLUTION TESTCOV
8 /CRITERIA=CIN(95) MXITER(5000) MXSTEP(50) SCORING(1)
9   SINGULAR(0.000000000001) HCONVERGE(0,
10  ABSOLUTE) LCONVERGE(0, ABSOLUTE)
11  PCONVERGE(0.000001, ABSOLUTE)
```

Fig. 4 SPSS syntax for Example 1

model discussed here. Most parts should be self-explanatory. The PREDICTORS statement (line 4) simply lists the factorial predictors and their interactions. Currently, covariates (interval-level predictors) cannot be included in DJMIXED. The NAME statement (line 7) is used to give the model a name. Names should be enclosed in quotes but can otherwise be freely chosen.

The second block of code, lines 8–10, prints out a summary table of the named model. The table is very similar to the one shown in Table 3. The output of the MIXEDMODEL subcommand (and the underlying SPSS MIXED command) is quite verbose, so this summary table can help users find the relevant numbers quickly.

To help the user keep track of the voluminous output, the first command has an extra option, OUTPUT. If this is set to "split" (the default), the full analysis results are directed to a secondary output window, and the model summary is automatically generated in the main output window. The other two possible values are "full" (no secondary window is used; the full SPSS output is shown without summary) and "none" (no output is generated at all).

Similar output can be generated with plain SPSS commands: The DJMIXED package prints out the equivalent syntax every time it is run. The plain SPSS syntax for all commands used here is listed in the online Appendix and, for this model, is shown in Fig. 4. When DJMIXED is not used, the output cannot be directed to two windows, and the model summary is not available (and neither is the model comparison that we will encounter later).

Example 2, stepwise analysis

The analysis presented above is a significant statistical improvement over the double approximation via F_1 and F_2 and the other approaches mentioned earlier. For most purposes, this analysis should suffice. For the interested reader, a statistically more thorough exploration of the significance of the priming × morph interaction can be made by presenting a stepwise analysis.

A stepwise mixed model analysis is very similar to a forward-stepping linear regression analysis: Starting with a very simple model, additional model terms are introduced until the point that model fit is no longer improved. A stepwise mixed model analysis should start with a model (often called the *null model*) that contains the random effects of participants and items but no other predictors (model 2, introduced below). In the next step, the fixed effects of priming and morph are added (model 3). With only two factors, the most complex model has two main effects and one interaction; this is model 1, discussed above. For each subsequent model, the improvement in model fit is evaluated against the cost of introducing extra factors or interactions. In an article that does not focus on statistical issues, not all steps have to be reported in full, but a summary of the steps taken to arrive at the final (best-fitting) model should be given.

In a stepwise regression analysis, we look at the R^2 to see whether the inclusion of additional terms improved the model fit of the data. In a mixed model approach, there is no direct equivalent of R^2, and the quantities Akaike information criterion (AIC) and -2LL are considered instead. These will be discussed in more detail below; for now, AIC can be viewed as an unstandardized, adjusted R^2, and -2LL will feed into a formal model comparison test, discussed below.

Mixed model 2

The second model fitted is a so-called null model, which is an intercept-only model without any predictors. In a mixed model analysis, the null model should contain the random factors as described above: adjustments to the intercept for

individual items and individual participants. The model formula is

$$Y_{pi} = \beta_0 + u_{0p} + u_{0i} + \varepsilon_{pi}$$

which indicates that the one RT obtained for each participant (p) and item (i) combination is modeled as the intercept β_0 with an adjustment to that intercept for the relative speed of this particular participant u_{0p}, an adjustment for the relative ease of this item u_{0i}, and residual error. Intercept adjustments u_{0s} and u_{0i} both sum to zero, as does ε_{ij}, such that the expected RT for each observation is the intercept, $E(Y) = \beta_0$.

The results of the null model are summarized in Table 4. The null model had four parameters and resulted in the following fit indices: -2LL = 25,443, AIC = 25,451. There were no fixed effects of interest. Both random effects were significant: for u_{0p}, $Z = 3.90$, $p = .000$; for u_{0i}, $Z = 5.31$, $p = .000$.

This short report on the null model does not include the fixed effect intercept and the variance explained by ε, since these effects do not aid in our understanding of the model. The reported fit indices will be used as a basis for comparison in the next steps. In an ANOVA context, the variances of u_{0p}, u_{0i}, and ε are summed to compute the contribution of each term to the total variance. In a mixed model context, this is not possible, because the u variances are usually correlated.

Mixed model 3

The third model extends the null model with the fixed factors priming and morph, but it does not include their interaction. No additional random effects are included in this model, but the existing random effects that adjust the intercept for participants and items may change due to the inclusion of the new predictors. The model formula is

$$Y_{pi} = \beta_0 + u_{0p} + u_{0i} + \beta_1 \text{Priming}_i + \beta_2 \text{Morph}_i + \varepsilon_{pi}$$

which indicates that the one RT obtained for each participant (p) and item (i) combination, is modeled as the intercept β_0, to which there is a participant-specific adjustment u_{0p} and an item-specific adjustment u_{0i}. The predicted RT varies by the factor priming, with a slope β_1, and by the factor morph, with a slope β_2. The word slope is used here for compatibility with the hierarchical linear modeling literature. The expected RT for each observation is $E(Y) = \beta_0 + \beta_1 \text{Priming} + \beta_2 \text{Morph}$.

The results of the third model are summarized in Table 5. The model had six parameters and resulted in the following fit indices: -2LL = 25,403, AIC = 25,415. The fixed effect of priming was highly significant, $F(1, 70) = 46.0$, $p = .000$, and so was the effect of morph, $F(1, 1012) = 5.39$, $p = .020$. Both random effects were significant: for u_{0p}, $Z = 3.90$, $p = .000$; for u_{0i}, $Z = 4.30$, $p = .000$. As compared with the null model (shown in Table 4), the variation related to participants did not change, whereas the variation related to items was reduced substantially. This is to be expected: The fixed factors morph and priming should explain some of the variation between items.

There are two ways to compare the fit of the null model (model 2) with the fit of this model (model 3). First, the values of AIC can be directly compared between the models, with lower values indicating a better fit. The AIC value for model 3 is 25,415, 36 points lower than the value of 25,451 obtained for model 2, indicating an improvement in fit. One cannot determine whether this difference is a significant improvement, since AIC values are unscaled. (However, as a rule of thumb, a difference of more than 10 points is usually an indicator of a significant improvement.)

A second way of comparing the models is via the likelihood ratio test (LRT). This test uses the raw fit measure *deviance* or *log-likelihood*. Because the likelihood value reported by most programs is log-transformed and multiplied by -2, the abbreviation used in SPSS and SAS is -2LL for *minus two times log-likelihood*. Similar to the F-test, which divides within-variance by between-variance, the LRT evaluates the relative fit of model 3 by dividing it by the fit of model 2. Division of two likelihoods is mathematically identical to the difference of two log-likelihoods, so we obtain LRT = 25,443 - 25,403 = 40 for the comparison between models 2 and 3 (the AIC is derived from the -2LL but also takes the number of parameters into

Table 4 Random effects for model 2, the null model

Model term		Random effects			
		Adjustment for	Variance	Z	p
u_{0p}	Intercept	Participants	5,129	3.90	.000
u_{0i}	Intercept	Words	3,507	5.31	.000
ε	Error	–	16,861	30.61	.000

Table 5 Fixed and random effects in the third model, main effects only

		Fixed effects			
Model term		Category	β	F	p
Priming		Unprimed	75.5	46.0	.000
Morph		Inflection	16.7	5.4	.020
		Random effects			
Model term		Adjustment for	Variance	Z	p
u_{0p}	Intercept	Participants	5,142	3.90	.000
u_{0i}	Intercept	Words	1,858	4.30	.000
ε	Error	–	16,891	30.6	.000

account, so the difference in AIC values is similar to the difference in -2LL).

The value of LRT can be statistically evaluated against a chi-squared distribution, using the difference in the number of model parameters as the degrees of freedom. Model 2 has four parameters, and model 3 has six, so a chi-squared with 2 degrees of freedom should be used. The test is $LRT(2) = 40, p < .0001$, which means that model 3 has a significantly better fit than model 2 (see also Table 6 for an overview of model comparisons).

Mixed model 1, revisited

Model 1, presented in Example 1 above, is similar to model 3, but model 1 also contains the interaction between morph and priming among the fixed effects. The random effects are still participant-specific and item-specific adjustments to the intercept.

The results of this model are summarized in Table 3 above. The model had seven parameters and resulted in -2LL = 25,399, AIC = 25,413. Model 1 has a lower (better) AIC value than does the previous models, although the difference with the third model (main effects only) is small and on the edge of significance ($p = .041$) according to the LRT.

Using this stepwise procedure, we can conclude that there is some statistical evidence for the presence of an interaction term morph$^{\times}$ priming: The interaction term is significant in the *F*-test presented in model 1, and model 1 is a slightly better fit of the data according to AIC values and the LRT test.

As compared with the normal ANOVA procedure, we have two statistical tests of the interaction at our disposal (*F*-test and LRT). Because both tests result in *p*-values that are rather close to our alpha level ($p = .042$ for *F*-test; $p = .041$ for LRT), it would be wise to investigate this interaction further in a follow-up experiment or by introducing additional predictors in the design (Keppel & Wickens, 2004), but for now, the conclusion of a statistically significant effect of the interaction can be maintained.

Figure 5 shows the syntax for the steps just taken. The null model (model 2) is specified by removing the PREDICTORS line (or by specifying PREDICTORS = NONE). After constructing model 3, the models are compared with each other via the COMPAREMODELS command. The output of these commands can be found (slightly reformatted) in Table 6.

In an article that does not focus on statistical issues, the detailed report of each modeling step given above can be reduced to the findings reported for the final model 1, including Table 3. The stepwise procedure can be summarized by including Table 6 and a short text such as the following: "A statistical model of the data was built from a null model (model 2) by stepwise adding all main effects (model 3), and all interactions (model 1). In each step, the more complex model showed a significant better fit of the data (see Table 6), leading to the final model 1."

Example 3: Contrasts and post hoc tests

The examples above have all dealt with one or more factors that each had only two levels (e.g., primed vs. unprimed). If a factor has more than two levels, a test for a significant effect of that factor is usually followed by an examination of which levels differ from each other. Similar to normal ANOVA procedures, this examination can be done with planned comparisons (also called *contrasts*) or omnibus comparisons (also called *post hoc tests*).

To illustrate this, the same data set is used as before, but the two factors priming and morph are now combined into one factor, form. Note that the factor form is constructed for didactic purposes only; this analysis will not clearly distinguish between primed and unprimed words, thereby obscuring one of the more important influences on RT.

Table 6 Overview of models 1–3, with the degrees of freedom (*df*), deviance (-2LL), and the Akaike fit index (AIC). Models are compared with the likelihood ratio test (LRT); a significant results indicates that the more complex model is preferable. For each comparison, base for comparison, the comparison *df*, and LRT value are shown. All models contain random adjustments to the intercept for participants and items (u_{0p} and u_{0i})

	Model details				Likelihood ratio		
	Fixed effects	*df*	-2LL	AIC	Comparison	*df*	LRT
2	None	4	25,443	25,403	–	–	–
3	Priming, morph	6	25,403	25,415	2	2	39.92***
1	Priming * morph	7	25,399	25,413	3	1	4.16*

*$p < .05$; **$p < .01$; ***$p < .001$

```
1  * syntax for the null model (Model 2).
2  DJMIXED /MIXEDMODEL
3              DV=rt
4              PPS=Participant
5              ITEMS=Word
6              NAME='null'.
7
8  DJMIXED /MIXEDMODEL
9              DV=rt
10             PREDICTORS = morph priming
11             PPS=Participant
12             ITEMS=Word
13             NAME='main effects'.
14
15 DJMIXED /COMPAREMODELS
16             NAME1='null'
17             NAME2='main effects'.
18
19 DJMIXED /COMPAREMODELS
20             NAME1='main effects'
21             NAME2='interaction'.
```

Fig. 5 DJMIXED syntax for remaining models and model comparison (line numbers added)

The new factor form has three levels—stem, inflected, and derived—matching the morphological status of the target word (see also Table 7).

Mixed model 4

This model has one theoretically relevant predictor, the factor form with three levels. The model contains an intercept and random effects for participant-specific and item-specific adjustments to the intercept.

The analysis yields a model with six parameters and fit indices -2LL = 25,395, AIC = 25,407. The fixed effect of form was highly significant, $F(2, 1945) = 90.7$, $p = .000$. The random effects adjusting the intercept were significant: for u_{0p}, $Z = 3.90$, $p = .000$;for u_{0i}, $Z = 3.11$, $p = .002$, so inclusion of all random effects was warranted.

Two planned comparisons were run, one comparing the levels inflection and derivation, and one comparing the average of these two levels with the stem form. Both planned comparisons were significant: Derivations versus inflections has a difference estimate of 39.56, $t(1949) = 4.62$, $p = .000$. Stems versus mean of inflections and derivations has a difference estimate of -75.43, $t(1942) = -12.80$, $p = .000$.

The DJMIXED syntax for this model is shown in Fig. 6. The specification of the fixed and random terms follows the same pattern as before. In line 7, post hoc tests are requested, which will be discussed below. Although the post hoc tests are a theory-free and cautious approach to determining any difference in levels, the application of planned comparisons is more popular. The drawback of using planned comparisons is that any application that is slightly data driven leads to highly inflated alpha rates. In other words, if the planned comparison was determined after obtaining the means (or preliminary means), the alpha rate is much higher than promised.

The DJMIXED syntax for planned comparison is shown in Fig. 6: Line 8 shows how the keyword CONTRAST is followed by the name of the variable, followed by a pipe symbol (|), followed by a specification of the contrast coefficients. Similar to standard ANOVA contrasts, there should be as many coefficients per contrast as there are levels of the variable. The coefficients in each contrast should sum to zero. The number of contrasts should equal the number of levels minus one, with individual contrasts separated by pipe symbols. As with the standard ANOVA, using independent or orthogonal contrasts is advisable (Keppel & Wickens, 2004).

A knowledge of the ordering of the levels of the variable is necessary to design and interpret contrasts. SPSS orders levels either numerically or alphabetically, depending on the values of the variable. It is advisable to use a numerical coding with value labels to avoid surprises. Here, numerical values are used, and the ordering is stem, derivation, and inflection. Note that while the post hoc output shows the variable labels, the output of planned comparison does not include this

Table 7 Relationship between the factors priming, morph, and form

Factor prime	Description	Factor morph	
		Inflection	Derivation
Unprimed	Presented first	Inflectional form	Derivational form
Primed	Presented following related inflection	Stem form	–
	Presented following related derivation	–	Stem form

```
1 DJMIXED /MIXEDMODEL
2      DV = rt
3      PREDICTORS = form
4      PPS = Participant
5      ITEMS = Base
6      MODEL = 'form as pred'
7      POSTHOC = form
8      CONTRAST = form | 0 1 -1 | 1 -0.5 -0.5  .
```

Fig. 6 Syntax for mixed model 4 with specification of post hoc tests and contrasts, line numbers added

convenience. The two planned comparisons are labeled L1 and L2 in the SPSS output.

Post hoc tests were requested in line 7; the option is followed by the name of the variable for which post hoc tests have to be computed. The additional SPSS output caused by this command has two parts: First, a table shows the mean and standard deviation for each condition, which is useful for creating graphs; second, six pairwise comparisons are performed that are all highly significant for the current data (p = .000 for each, significant after Sidak adjustment for multiple comparison).

Because post hoc tests involve multiple comparisons, the familywise alpha has to be controlled. Instead of the familiar Bonferonni *approximation* to the correct alpha for multiple comparisons, the exact formula for alpha correction, as proposed by Šidák, is recommended (see also Abdi, 2007).

Note that all comparisons are based on expected means, not observed means. This implies that a comparison based on a model that does not fit the data well may result in unreliable post hoc comparisons.

Regression diagnostics and transforms

In both regression and the ANOVA, the distribution of the residuals can inform us about the overall fit of the data and about the specifics of the fit. The DJMIXED package can produce four informative plots: a histogram of residuals, a Q-Q plot of observed versus expected residuals, a detrended Q-Q plot, and a plot of normalized residuals by predicted values. We will look only at the first plot here.

The histogram of residuals for Model 1 is shown in the left panel of Fig. 7. This distribution should be close to normal, and some appreciable differences exist for this model (the Kolmogorov–Smirnov test is significant, KS (2004) = 0.106, p = .000, indicating a significant difference between the observed curve and a normal curve).

On the one hand, the aim of psycholinguistic studies is usually not to provide a perfect fitting model of the data but to determine whether certain factors make a significant contribution to the prediction of RT or not. Under that view, a moderate to small departure from normality should not overly worry us, although it should be reported. However, ill-fitting residuals can be a sign of a model that does not capture the data very well. The significance of factors and their interactions may be hidden by or caused by the fact that the model does not fit well (see Rouder, Tuerlinckx, Speckman, Lu, & Gomez, 2008, for a promising approach to increasing the model fit of RT data).

A commonly suggested transform for RT data is the logarithmic transform (Keppel & Wickens, 2004; Van Breukelen, 2005), which will reshape the distribution of

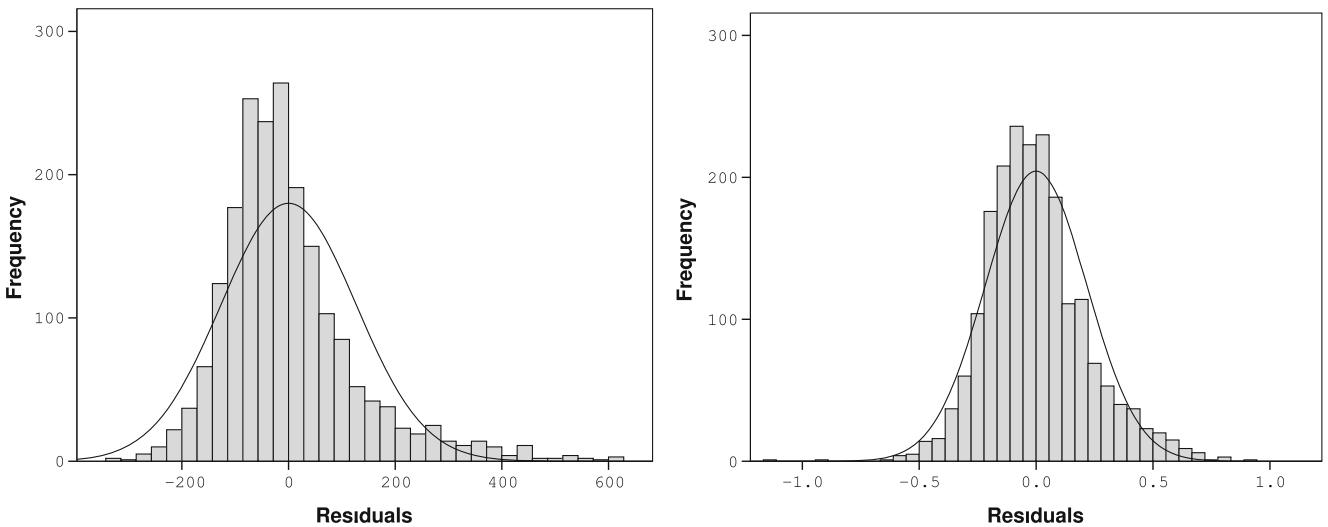

Fig. 7 Histogram of residuals (with normal curve superimposed) for models with the dependent variable reaction time [RT; left panel] and log-transformed RT [$log(RT–100)$; right panel]

RT data so that its heavy right tail is removed and it becomes more akin to a normal distribution. A mixed model analysis of log-transformed RTs resulted in the same significance levels as in model 1 but improved the distribution of residuals (Fig. 7, right panel).

Whether the additional fit gained from log-transformation is important should be decided on a case-by-case basis. Log transforms are not frequently used in the psychology of language literature, but they are common in neighboring fields such as cognitive modeling and corpus research. One issue that arises is that a simple log transform [$x_l = log(x)$] will often turn the shortest RTs into outliers. The solution is to subtract an estimate of the minimum RT, effectively moving the zero point to the right (Rouder et al., 2008): For the log-RT analysis reported in Fig. 7, $x_l = log(x - 100)$ was chosen.

Raw data plots and plots of residuals after intermediary models are fit can also be very instructive as to the structure of the data set. Textbooks such as Raudenbush and Bryk (2002) and Piñheiro and Bates (2000) show many examples of this. For the present data set, a plot of the observed data for the primed versus unprimed condition (for derivational pairs only) is shown in Fig. 8. In this plot, every item set is represented by a single line. Evidence for word-specific adjustments to the intercept can be found at the left edge of the figure, which shows that the item-specific intercepts differ substantially. The statistically significant but not completely compelling interaction effect morph × priming may well be due to the heterogeneity of the priming effect on the items: In the figure, three items show negative priming effects (dashed heavy lines), and six items show much stronger effects of priming than do the others (solid heavy lines). This could be unsystematic variation of the efficacy of priming, but facing such data, it is wise to investigate whether there are any factors that may help explain this.

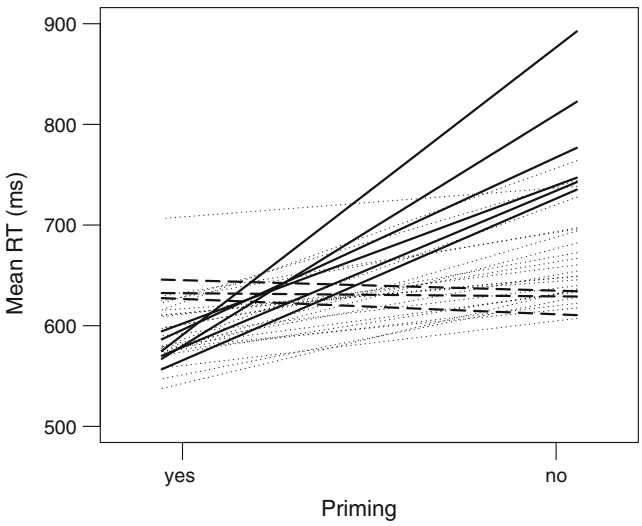

Fig. 8 Plot of observed primed and unprimed reaction times for each derivational pair, with negative priming (dashed lines) and strong positive priming (solid lines) highlighted

More complex models

In almost every case, the stepwise analysis presented above should be sufficient to draw psycholinguistically valid conclusions from the data. The mixed model can be further extended in two ways, which will be briefly outlined here.

The first extension concerns covariates: On some occasions, there are known covariates that may help explain the differences between items (or more rarely, participants). If there are theoretical reasons to expect that, for example, log frequency will co-determine RTs, this predictor should be included in the model. The mixed models described here are very similar to normal regression models, and an effect of frequency can be added as a predictor in a straightforward way. The model formula and the DJMIXED syntax are included in the online Appendix.

The second possible extension concerns the way differences between items and participants influence the expected outcomes. In the models so far, the random effect of items (and participants) has been added to the intercept to indicate whether an item is *generally* easy or hard. It is possible that the effect of a predictor (say, priming) also differs between items (or between participants). One way to account for that is to include a second random effect for items, which modifies the strength of the effect of priming. Mathematically, a new random effect u_{1i} is added to the β for priming, as shown in this partial formula: $Y_{ij} = ... + (\beta_1 + u_{1i}) \cdot \text{PrimeType}_i + ...$ In mixed model parlance, the random effect u_{1i} modifies the slope (β) of priming.

A number of statistical complications arise with this type of analysis, and the online Appendix goes into some detail on how to work around these. However, there are further reasons why this type of analysis may not be applicable to most psycholinguistic experiments.

First of all, the exact structure of the random effects is rarely a psycholinguistic goal in itself. A model with a random effect on the slope of priming does not give a better theoretical explanation; it merely adds a device for capturing unexplained variance. A well-chosen covariate is often a better option, since it does add theoretical strength to the model.

Second, extracting three or more random effects from the data is demanding. Psycholinguistic data tend to have one observation per participant–item combination, and the numbers of items and participants tested are sizable, but not in the hundreds. Both of these factors limit the "carrying capacity" (Nezlek, 2008) of the data. The two random effects related to items (u_{0i} and u_{1i}) are most often

correlated, which makes it harder to arrive at estimates for each of them, necessitating large number of items and participants.

Third, the extended analysis assumes that the item-related random effects modifying the intercept and the slope of priming are independent influences, which may be correlated. As has been argued by Rouder et al. (2008), faster items often show less variability and are, therefore, inherently less sensitive to priming. In a sophisticated model that creates a connection between an item's mean and its standard deviation, these authors were able to show that there was no need for item- or participant-specific adjustment to slopes once the correlation between mean and standard deviation was taken into account.

In sum, there seem to be few compelling reasons to add random effects that modify slopes. For psycholinguistic data, models like those presented above already provide a better description of the data than do classical ANOVA models, and it may well turn out that the extra complications caused by adding random effects modifying slope are rarely necessary in practice. Authors should run the usual regression diagnostics to determine whether the data were fitted reasonably well or whether further statistical explorations are necessary.

Discussion

This article has presented a simple framework for addressing the issue of random participants and random items in language experiments. The DJMIXED extension to SPSS should put this mixed models analysis within the reach of every psycholinguist. It was argued that the results of mixed models are easy to interpret, while staying much closer to the data than do the other approaches that are currently in common use (min F', F_1/F_2, treating items as fixed).

Mixed models should be used only when the data set is large enough and after outliers and wrongly coded observations have been removed. Conceptually, a separate regression line is estimated for each level of participants and also for each level of items, so an extreme outlier can have a large influence if the number of observations per participant or the number of observations per item is low. As compared with an ANOVA, the restrictions on the data imposed by mixed modeling are very relaxed, since missing data and unequal cell sizes are not a problem and homoscedascity is not an a priori requirement either. Mixed models require equality of residual variance; that is, the predictors should capture not only the difference in average RTs, but also any difference in variability of RTs. For most data sets, this seems a tenable assumption (but see Rouder et al., 2008), and there are currently few alternatives for those cases in which this assumption is mildly violated.

Mixed models are a relatively recent extension to the statistical canon, and although the pace of development has slowed down, further improvements to these models and their evaluation will most certainly be found. However, the methods of model evaluation that are suggested here (F-tests and LRT) have shown their merits outside of mixed modeling, and they are implemented in major statistical packages such as SPSS and SAS and are generally recommended in various fields of science.

Like most statistical tests, these tests are not perfect under all circumstances: Using the LRT to test for the inclusion of random effects is slightly conservative when the difference between parameters in the two models is used as the degrees of freedom (Kreft & De Leeuw, 1998; Stram & Lee, 1994). The alternative of using a 50/50 chi-squared mixture was suggested by Stram and Lee and is adopted in the Appendix to this article and elsewhere (Kreft & De Leeuw, 1998; Verbeke & Molenberghs, 2001). But criticisms against this procedure have been leveled (Baayen, Davidson and Bates 2008; Piñheiro & Bates, 2000, p. 70), suggesting that it may still be slightly conservative (not rendering enough significant results). Faraway (2006) suggested using a parametric bootstrap (cf. Janssen, Bickel & Zúñiga 2006) to correct the p-values of the LRT when testing for the inclusion of random effects, and this procedure has the advantage over the solution suggested by Baayen et al. (2008) of not depending on a Bayesian framework.

LRT tests for the inclusion of fixed effects are also widely used (Kreft & De Leeuw, 1998; Raudenbush & Bryk, 2002; Verbeke & Molenberghs, 2001), but some have argued that this test may be too liberal (allowing too many significant results) for tests on certain data sets (Baayen et al., 2008; Piñheiro & Bates, 2000, p. 88). The example given by Piñheiro and Bates cautions the reader not to test for the inclusion of fixed effects with a very large number of levels, as compared with the total number of observations. It is shown that the LRT can be too liberal when testing for the inclusion of a factor with 15 levels in a data set with only 60 observations. In practical psycholinguistic applications, the number of levels of a fixed factor hardly ever exceeds five, so if the general recommendations for sizable numbers of participants and items—and, therefore, observations—are followed, this criticism should not overly concern us. Raudenbush and Bryk evaluated the merits of the LRT, as compared with a multiple comparison procedure similar to the contrasts discussed above, and concluded that the LRT is a valid procedure that will produce results nearly identical to those of multiple comparison (Raudenbush & Bryk, 2002, p. 61), while the LRT is much easier to implement. When models for the inclusion of a fixed effect are compared, the multiple comparison tests are similar to the F-test that was used here in conjunction with the LRT.

SPSS and SAS report z-tests on individual model βs, which should not be used for drawing conclusions about the importance of predictors. The z tests can be very conservative, and Raudenbush and Bryk (2002) suggested using a t-distribution instead. The issue was side-stepped here by using F-tests (technically, Type 3 F-tests) of fixed effects instead. In this article, values of β were reported in tables to offer the reader an insight into the direction and magnitude of the effect, but the p-values listed are derived from the F-tests on the fixed effects in the analysis. The advantage of the omnibus F-tests is that they are available in SPSS and in SAS and that they are similar in interpretation to the normal ANOVA tests. The F-tests also produce one significance value for factors with more than two levels, whereas multiple significances result if the z-tests are followed (one z-test is presented for each β). This technique is followed by almost every text on HLM and mixed modeling (Faraway, 2006; Hox, 1995; Kreft & De Leeuw, 1998; Raudenbush & Bryk, 2002; Singer, 1998; Snijders & Bosker, 1999; Verbeke & Molenberghs, 2001).

F-tests (including Type 3 F-tests) for normal ANOVAs and mixed-models alike have been criticized by Venables (1998). To evaluate the significance of fixed effects, it has been suggested to askew F-tests (Baayen, 2008; Bates, 2006, 2008) and use MCMC (Monte Carlo Markov chains), a simulation technique based on Bayesian principles to approximate the significance of each fixed effect on an analysis-by-analysis basis. This technique has certain theoretical advantages for data with smaller numbers of cases, but it is not implemented in SPSS or SAS. It also requires one to work within a Bayesian inference framework, which has various advantages and disadvantages that fall outside of the scope of this article. In a discussion of which test to use, Faraway (2006, 2009) recommended the combined use of the F-test and the LRT.

Of course, statistics is a scientific discipline just like psycholinguistics, and dissenting opinions, alternative approaches. and progressing insights are par for the course. Mixed models, and hierarchical linear models as their special case, are a mature technique, and they have been implemented in the major statistical packages since 1996 (SAS), 2000 (R), and 2002 (SPSS). Straightforward and relatively uncomplicated applications of mixed models, such as advocated in this article, are used in biology (O'Connor, Bruno, Gaines, Halpern, Lester, Kinlan & Weiss 2007), educational research (Raudenbush & Bryk, 2002), social psychology and personality research (Nezlek, 2008), signal detection theory (Rouder & Lu, 2005), and many other fields. Mixed models are easy to construct in SPSS and SAS, and the mixed model results are straightforward to understand when the focus remains on the fixed effects. It is time for psycholinguistics to leave the realm of F_1/F_2 testing and move to mixed modeling as a standard means of assessing significance.

Acknowledgements I would like to thank Nicolas Dumay and Lea Hald for constructive comments on an earlier draft. Thanks to my former colleagues at the University of Kent (and especially Joachim Stoeber) for their continuing support. This study has been shaped by a number of workshops on mixed modeling that I have given; thanks to all participants for clarifying to me what makes mixed modeling so hard to understand.

References

Abdi, H. (2007). The Bonferonni and Šidák corrections for multiple comparisons. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 103–107). Thousand Oaks, CA: Sage.

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

Baayen, R. H., Davidson, D. J., & Bates, D. (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,* 55–65.

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

Bates, D. (2006). Fitting linear mixed models in R. *R News, 5,* 27–30.

Bates, D. (2008). *The lme4 package* [Computer software manual]. Retrieved from http://cran.r-project.org/web/packages/lme4/lme4.pdf

Cheng, C.-P., Sheu, C.-F., & Yen, N.-S. (2009). A mixed-effects expectancy-valence model for the Iowa gambling task. *Behavior Research Methods, 41,* 657–663.

Clark, H. H. (1973). The language-as-a-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.

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. London: Chapman & Hall/CRC.

Faraway, J. J. (2009). *Changes to the Mixed Effects Models chapters in ELM* [Online update to the book by Faraway, 2006]. Retrieved from http://www.maths.bath.ac.uk/jjf23/ELM/mixchange.pdf

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

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

Jackson, S., & Brashers, D. E. (1994). *Random factors in ANOVA*. Thousand Oaks, CA: Sage.

Janssen, D. P., Bickel, B., & Zúñiga, F. (2006). Randomisation test in language typology. *Linguistic Typology, 10,* 419–440.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive Science, 32,* 1403–1424.

Maxwell, S. E., & Bray, J. H. (1986). Robustness of the quasi F statistic to violations of sphericity. *Psychological Bulletin, 99,* 416–421.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11,* 81–91.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2,* 842–860.

O'Connor, M. I., Bruno, J. F., Gaines, S. D., Halpern, B. S., Lester, S. E., Kinlan, B. P., et al. (2007). Temperature control of larval dispersal and the implications for marine ecology, evolution, and conservation. *Proceedings of the National Academy of Sciences, 104,* 1266–1271.

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

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

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

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 194,* 337–350.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18,* 4.

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,* 221–250.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12,* 573–604.

Rouder, J. N., Tuerlinckx, F., Speckman, P., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review, 15,* 1201–1208.

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

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248–1284.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *24*, 323–355. Retrieved from http://gseweb.harvard.edu/%7Efaculty/singer/Papers/Using%20Proc%20Mixed.pdf

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics, 50,* 1171–1177.

Van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70,* 359–376.

Venables, W. N. (1998, October). *Exegeses on linear models*. Paper presented to the S-PLUS User's Conference. Washington, DC. Retrieved from http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf on 18 Nov 2008.

Verbeke, G., & Molenberghs, G. (2001). *Linear mixed models for longitudinal data*. New York: Springer.

Wickens, T. D., & Keppel, G. (1983). On the choice of design and of test statistic in the analysis of experiments with sampled materials. *Journal of Verbal Behavior and Verbal Behavior, 20,* 296–309.

Wike, E. L., & Church, J. D. (1976). Comments on Clark's "The language-as-fixed-effect fallacy. *Journal of Verbal Learning and Verbal Behavior, 15,* 249–255.